Normandie Université

Habilitation à Diriger des Recherches

Pour obtenir le diplôme d'habilitation à diriger des recherches

Spécialité INFORMATIQUE

Préparée au sein de l'Université de Caen, Normandie

Rôle de la sémantique morpho-dispositionnelle pour l'accès non visuel aux documents numériques

Présentée et soutenue par
Fabrice MAUREL

HDR soutenue publiquement le 13 juin 2022 devant le jury composé de		
Jean-Yves Antoine	PU, LIFAT, Université de Tours	Rapporteur
Antoine Doucet	PU, L3i, Université de La Rochelle	Rapporteur
Olivier Gapenne	PU, COSTECH, Université de Technologie de Compiègne	Rapporteur
Julie Lemarié	PU, CLLE, Université de Toulouse 2	Examinatrice
Mustapha Mojahid	MCF HDR, IRIT, Université de Toulouse 3	Examinateur
Gaël Dias	PU, GREYC, Université de Caen Normandie	Examinateur (Directeur de l'HDR)
Stéphane Ferrari	MCF HDR, CRISCO, Université de Caen Normandie	Examinateur

UNICAEN
UNIVERSITÉ
CAEN
NORMANDIE

GREYC

Table des matières

Introduction et mise en perspective

Une Habilitation à Diriger des Recherches ?

Vient un moment de la carrière d'enseignant chercheur où se fait ressentir plus que d'habitude le besoin d'un bilan à mi-parcours ; de s'extraire des impératifs du quotidien et de ses obligations professionnelles variées et tumultueuses, pour questionner avec un peu plus de recul les spécificités de son évolution scientifique ; en bref, prioriser l'important sur l'urgent. C'est ce temps nécessaire d'introspection que j'ai décidé de faire coïncider avec la rédaction de mon habilitation à diriger des recherches. Cette concomitance a contribué à guider mes choix pour l'organisation de ce document puisqu'il s'est agi de mettre en cohérence le travail effectué depuis mon doctorat, obtenu en 2005 ; de suivre ce fil directeur et ses intrications avec mes objectifs actuels ; et envisager les nouvelles ramifications qui vont orienter mon activité à court et moyen terme.

Un deuxième élément fondera mes choix de présentation. Il concerne les différents projets financés qui émaillent cette chronologie. Ces derniers permettront d'expliquer et de valoriser les différentes collaborations pluridisciplinaires mises en place, en particulier dans la réciprocité des apports entre disciplines dites « dures » et « humaines » (notons que, non convaincus par les deux qualificatifs exploités ici, certains pourraient ironiser en opposant à ces termes respectivement sciences « molles » et « inhumaines » !).

Outre ce « qui ? » et ce « quoi ? », le troisième principe que j'espère dégager par ce travail de rédaction est un regard original sur « comment » emporter le défi de l'accessibilité numérique qui implique, entre autres, les domaines informatiques du Traitement Automatique des Langues (TAL) et de l'Interaction Homme Machine (IHM). Cette réflexion porte d'une part sur l'attention à donner à l'usage des interfaces et à la maximisation de la capacité d'autodétermination de l'usager ; d'autre part sur le peu d'occasions de partage des recherches, pourtant par nature nombreuses, à l'intersection de ces deux communautés scientifiques.

Cette mise en perspective justifie une organisation du manuscrit en trois parties, alimentées selon le plan détaillé ci-après.

D'où viens-je ?
PARTIE I : *Layout*, Multimodalité et Handicap

Après une formation orientée dès mon Diplôme d'Études Approfondies (DEA) à l'Université Paul Sabatier de Toulouse sur l'informatique de l'image et du langage, et en particulier sur le

traitement automatique de la métonymie, j'ai choisi d'intégrer l'équipe de recherche *Dialogue InterAction Multimodalité Accessibilité et Nouvelles Technologies* (DIAMANT) de l'*Institut de Recherche en Informatique de Toulouse* (IRIT) dont les spécificités ont forgé durablement mes centres d'intérêt scientifiques ; elles sont en cela les premiers brins qui constituent le fil directeur de mon activité de recherche.

Une spécificité déterminante concerne la problématique de la fracture numérique et du handicap. Cet aspect était à la fois au cœur des questionnements scientifiques de l'équipe mais également inhérent à la composition même de celle-ci. Avoir travaillé sur ces thèmes en collaboration avec des personnels à la fois ingénieurs ou chercheurs en situation de handicap et usager des solutions que nous étudions, m'a convaincu de la nécessité d'une démarche à la fois tournée vers l'utilisateur et participative.

La deuxième spécificité de mon cadre de travail a été la démarche fortement pluridisciplinaire imposée au travers d'un sujet de thèse inscrit dans le programme « cognitique » du CNRS. J'ai pu apprécier l'intérêt d'un co-encadrement [1] suscitant la confrontation des angles de vues des psychologues, linguistes et informaticiens pour apporter des réponses à des questions qui n'auraient pu être résolues par une discipline seule.

La troisième spécificité résidait dans le sujet même du projet, nouveau et original, dans lequel s'inscrivait le travail de thèse qui m'était demandé pour faciliter l'accès non visuel aux documents numériques structurés : comment intégrer dans les systèmes de synthèse de la parole à partir de texte, qui s'appuient sur la phrase comme unité d'analyse, la prise en compte de la mise en page/forme qui peut relever d'un niveau de traitement intra- ou supra-phrastique ? La dureté de ce verrou, à la fois technologique, scientifique, et toujours d'actualité, m'a conduit à m'interroger sur la notion de modalités et de substitutions sensorielles ou encore de sémantique de la mise en forme des documents numériques, dont les tenants et les aboutissants m'animent encore aujourd'hui.

Je présenterai dans cette partie les éléments qui permettent de comprendre mes positions actuelles sur ces sujets et les projets qui en portent l'empreinte. Il s'agira de développer les points suivants :

- Transmodalité et multimodalité comme objets de recherche ;
- Étude du cas de l'écrit vers oral : sémantique morpho-dispositionnelle des textes et application à l'accès non visuel aux documents numériques ;
- Évaluation des premiers modèles et premières limites ;
- Dichotomie page / Image de Page et accès non-visuel interactif aux textes.

1. par Jean-Luc Nespoulous en neuropsycholinguistique et Nadine Vigouroux en informatique et interaction homme-machine

Où suis-je ?
PARTIE II : Stratégies de lecture non visuelle, rapide, globale, interactive

Les expérimentations avec les non-voyants durant mon travail de thèse ont mis en évidence les limites d'un modèle basé sur la reformulation pour conserver à la fois toute la sémantique véhiculée par la mise en forme mais également les capacités de compréhension et de mémorisation qu'elle participe à faciliter. Une des perspectives était de considérer un même document selon deux partis pris distincts : celui d'un contenu articulable logico-thématiquement structuré mais également celui de l'image d'une page visuellement contrastée. Cette dernière qualité participerait fortement à l'activation d'une capacité de vision globale et au développement de stratégies de lecture de haut niveau. Une fois cette dichotomie page/image de page acceptée, si le premier parti pris reste compatible avec des stratégies orales de reformulation, l'autre peut s'appuyer sur un accès non visuel au travers de la modalité tactile. C'est de cette idée qu'est né mon premier projet soutenu par l'Agence Nationale de la Recherche (ANR) en 2013, en tant que coordinateur scientifique : Accès par Retour Tactilo-oral Aux Documents Numériques (ART-ADN).

Les récentes conclusions de ce projet mettent également en évidence, dans les conditions expérimentales que nous avons pu mener, un certain nombre de limites. L'accès tactile à la mise en forme permettrait de développer des stratégies intéressantes pour la recherche rapide d'éléments (*scanning*) dans un document mais insuffisantes pour développer celles qui s'appuieraient sur une vision globale et rapide (*skimming*). Après deux Contrats Plan Etat Région (CPER) d'évaluation de sa faisabilité, nous avons développé un nouveau projet, TAGTHUNDER, qui a trouvé un financement par le Fonds national pour la Société Numérique (FSN), géré par la Banque Publique d'Investissement (BPI) dans le cadre de l'appel « Accessibilité numérique » du Plan Investissement Avenir 2 (PIA2). L'objectif consiste à faire retrouver une telle aptitude non visuelle de « premier regard » en construisant des versions sonores spatialisées de pages Web intégrées dans des systèmes interactifs appropriés.

Je détaillerai les idées qui ont jalonné ce cheminement scientifique à travers les points suivants :

- Herméneutique, énaction et interprétation des textes : vers l'étude de systèmes de substitution sensorielle automatique ;
- Les 5 piliers de ma démarche de recherche ;
- Application au projet TactiNET : métaphore de la canne blanche ;
- Application au projet TagThunder : métaphore de la *cocktail party* ;
- Résultats autour de la segmentation automatique de pages Web

Où vais-je ?
Perspectives et conclusion

Après avoir survolé dans les deux parties précédentes 20 années de travaux scientifiques, et leur convergence vers mes problématiques actuelles, il reste à explorer la manière dont ils soutiendront la cohérence des futurs projets.

Mes centres d'intérêt s'inscriront transversalement dans trois cadres de travail qui circonscriront mes actions scientifiques et l'organisation de cette partie conclusive :

- un cadre applicatif autour de la e-santé et du handicap sensoriel ou cognitif (non-voyance, pathologies psychiatriques ou sociales) dans des tâches spécifiques (lecture, recherche d'information) ;
- un cadre collaboratif pluridisciplinaire (informatique de l'image et du langage, linguistique, psychologie, sciences sociales, neurosciences) orienté à l'intersection du traitement automatique des langues et de l'interaction homme machine ;
- un cadre de recherche en intelligence artificielle, tourné vers la combinaison d'approches statistiques plus contrôlables, et la transposition, sans pertes informationnelles ou cognitives, d'une ou plusieurs modalités de présentation vers une ou plusieurs autres.

Ces trois cadres de travail me permettrons de développer l'avenir scientifique, technique et industriel des projets TactiNET et TAGTHUNDER et, au-delà, mes questionnements de recherche à court et moyen terme qui sont susceptibles d'en découler :

- Comment exploiter les documents dans toutes leurs dimensions ? Ou la question de la sémantique morpho-dispositionnelle appliquée à la complexité des pages Web ;
- Comment découper une page Web en zones d'intérêt de lecture ? Ou la question de l'accès global aux documents ;
- Comment représenter une zone d'intérêt de lecture par ses points saillants ? Ou la question de l'accès local aux documents ;
- Quelles influences mutuelles entre les relations rhétoriques des zones de lecture et les relations sémantiques des contenus ? Ou les questions du rapport texte/image et de l'inclusion sémantique des unités lexicales ;
- Comment transposer efficacement une structure visuelle complexe en paysage tactile et/ou sonore ? Ou la question de la définition des contours d'une " théorie gestaltiste non visuelle " ;
- Comment interagir de manière autonome à la fois localement et globalement avec les spécificités perceptives de " micromondes " tactiles et/ou sonores ? Ou les questions de l'accessibilité numérique et des interfaces langagières énactives.

Aussi, je conclurai en décrivant comment les recherches que je souhaite diriger traverseront ces perspectives par les nouveaux projets et encadrements déjà engagés.

Première partie

Layout, Multimodalité et Handicap

Contexte et environnement :

- *Institut de Recherche en Informatique de Toulouse (IRIT)*
- *Équipe Dialogue InterAction Multimodalité Accessibilité et Nouvelles Technologies (DIAMANT)*
- *Doctorant et ATER - publications 2001-2006*

Après une formation orientée dès mon Diplôme d'Études Approfondies (DEA) à l'Université Paul Sabatier de Toulouse sur l'informatique de l'image et du langage, et en particulier sur le traitement automatique de la métonymie, j'ai choisi d'intégrer l'équipe de recherche *Dialogue InterAction Multimodalité Accessibilité et Nouvelles Technologies* (DIAMANT) de l'*Institut de Recherche en Informatique de Toulouse* (IRIT) dont les spécificités ont forgé durablement mes centres d'intérêt scientifiques ; elles sont en cela les premiers brins qui constituent le fil directeur de mon activité de recherche.

Une spécificité déterminante concerne la problématique de la fracture numérique et du handicap. Cet aspect était à la fois au cœur des questionnements scientifiques de l'équipe mais également inhérent à la composition même de celle-ci. Avoir travaillé sur ces thèmes en collaboration avec des personnels à la fois ingénieurs ou chercheurs en situation de handicap et usager des solutions que nous étudions, m'a convaincu de la nécessité d'une démarche à la fois tournée vers l'utilisateur et participative.

La deuxième est la démarche fortement pluridisciplinaire qui m'a été imposée au travers d'un sujet de thèse inscrit dans le programme « cognitique » du CNRS. J'ai pu apprécier l'intérêt de la confrontation des angles de vues des psychologues, linguistes et informaticiens pour apporter des réponses à des questions qui n'auraient pu être résolues par une discipline seule.

La troisième réside dans le sujet même du projet, nouveau et original, dans lequel s'inscrivait le travail de thèse qui m'était demandé pour faciliter l'accès non visuel aux documents numériques structurés : comment intégrer dans les systèmes de synthèse de la parole à partir de texte, qui s'appuient sur la phrase comme unité d'analyse, la prise en compte de la mise en page/forme qui peut relever d'un niveau de traitement intra- ou supra-phrastique ? La dureté de ce verrou, à la fois technologique, scientifique, et toujours d'actualité, m'a conduit à m'interroger sur la notion de modalités et de substitutions sensorielles ou encore de sémantique de la mise en forme des documents numériques, dont les tenants et les aboutissants m'animent encore aujourd'hui.

Je présenterai dans cette partie les éléments qui permettent de comprendre mes positions actuelles sur ces sujets et les projets qui en portent l'empreinte. Il s'agira de développer les points suivants :

- Transmodalité et multimodalité comme objets de recherche ;
- Étude du cas de l'écrit vers oral : sémantique morpho-dispositionnelle des textes et application à l'accès non visuel aux documents numériques ;
- Évaluation des premiers modèles et premières limites ;
- Dichotomie page / Image de Page et accès non-visuel interactif aux textes.

1. Transmodalité et multimodalité comme objets de recherche

La communication humaine est par nature multicanale [148]. Cette simple remarque soulève au moins deux questions importantes, relatives à la manière dont nous optimisons cette caractéristique, consciemment ou non, pour aller dans le sens dicté par nos intentions communicatives :

- comment sont traitées les informations qui transitent simultanément par plusieurs canaux sensorimoteurs, que ce soit pour la production de messages (geste, oral, écrit…) ou pour la perception d'informations (auditives, visuelles, tactiles…) ? Cette question pose la problématique d'une gestion multimodale de la communication ;
- comment pallier une situation de communication dégradée par la défectuosité, temporaire ou permanente, d'un canal sensorimoteur particulier ? Cette question pose la problématique de la transmodalité, c'est à dire de la transposition de l'information d'une modalité vers une autre.

1.1. La transmodalité pour étudier la multimodalité

Le travail qui a conduit à la rédaction de ma thèse abordait ces questions au travers de l'étude de la présentation « transmodale » et multimodale d'un document électronique. La question est d'autant plus sensible aujourd'hui que les technologies sont de plus en plus transparentes et « intelligentes », que les supports de lecture et les pratiques se sont encore diversifiés, et que l'information continue d'être toujours plus foisonnante, composite et multisource. Le problème de leur consultation et de leur accessibilité sur de nouvelles interfaces, afin de pallier les difficultés (handicap, mobilité, taille d'écran, …), s'en voit même accru dans certaines circonstances et la fracture numérique ouverte n'a pas encore connue la véritable réduction espérée. Les interfaces et les moyens d'interaction proposés sont-ils efficaces, faciles d'utilisation, acceptés par tous et dans toutes les situations ? Sont-ils par exemple capables de proposer une présentation de l'information qui favorise notre gestion multimodale naturelle de l'information ou, le cas échéant, de transformer une information depuis une modalité dans une autre afin de la rendre plus accessible ? Doit-on préférer une approche de ces questions plutôt hypothético-déductive ou enpirico-inductive ?

Lorsque la multimodalité est abordée dans le cadre de la communication de la machine vers l'homme, les modèles proposés ont le plus souvent pour objet l'architecture fonctionnelle d'un système multimodal ou l'inventaire exhaustif des possibilités qui peuvent être offertes à l'utilisateur. D'autres systèmes développés peuvent reposer sur une modélisation empirico-déductive

en fonction de critères d'usage ou de résultats d'évaluations expérimentales exploratoires. Bien qu'il soit possible d'analyser l'utilisabilité du système au regard du pouvoir d'expression des modalités impliquées [116], les études ont montré assez vite qu'il était difficile de faire un lien entre les possibilités offertes par le système interactif du point de vue de la multimodalité et de l'usage qui en est fait[25] : ce n'est pas parce qu'un système interactif offre un type de multimodalité qu'elle sera forcément exploitée par l'utilisateur ; autrement dit, l'utilisabilité n'implique pas l'utilité[157]. Aussi, la richesse expressive des différentes modalités et la complexité des relations qu'elles entretiennent entre elles, rendent bien difficile une modélisation hypothético-déductive des conditions d'efficacité d'une présentation multimodale de l'information. Une telle modélisation permettrait pourtant de restreindre substantiellement en amont le champ des investigations expérimentales. Que le système multimodal à concevoir veuille intégrer différentes modalités selon un schéma d'interaction programmé ou préétabli (notion d'adaptabilité des interfaces), ou qu'il choisisse dynamiquement la manière d'utiliser les modalités en fonction du contexte d'interaction (notion d'adaptativité des interfaces) [176], il nous est apparu qu'il convenait dans un premier temps de se pencher sur les caractéristiques de chaque modalité de présentation à la fois du point de vue de leur pouvoir d'expression et de leur traitement cognitif.

La méthode souvent utilisée pour poser un tel cadre repose sur l'usage ou sur l'élaboration de taxonomies dont l'intérêt principal est de caractériser les modalités à travers des distinctions jugées utiles pour diriger la conception d'interfaces multimodales (au niveau du choix des modalités et de la compatibilité de leur combinaison). L'analyse de l'état de l'art [15, 17, 26, 40, 55, 98, 115] montrait que ces différentes méthodes de conception ne fournissent pas d'éléments clairs pour identifier des règles de choix et de composition des modalités qui assurent une utilité de l'interface. Le travail réalisé s'est donc appuyé sur une approche originale différente basée sur la notion de transmodalité. Ce terme caractérise la conversion d'une modalité de communication vers un canal de communication lié à un des cinq sens du récepteur. Nous avons envisagé cette notion d'un point de vue méthodologique pour offrir un cadre hypothético-déductif argumenté à une multimodalité utile et efficace :

- dans un premier temps, l'utilité de la multimodalité se mesure proportionnellement à notre incapacité à transposer sans « pertes » une information d'une modalité originelle vers une modalité a priori plus adaptée à une situation de communication particulière ;
- dans un second temps, l'efficacité de la multimodalité se mesure proportionnellement à sa capacité de réponse aux problèmes identifiés dans le point précédent.

1.2. Application à la transposition automatique à l'oral

Il est parfois fait l'hypothèse que la lecture dite silencieuse a eu un développement relativement récent. Certains ont vu des raisons technologiques au déplacement d'une lecture orale vers une lecture silencieuse ; l'introduction de l'imprimerie aurait conduit à un changement radical d'une culture orale à une culture visuelle, d'une lecture orale en groupe à une lecture silencieuse individuelle [107, 118]. Pourtant la littérature indique [30] que la notion de société médiévale purement orale est une simplification puisque des agencements de pages non linéaires

complexes étaient courants. De nombreuses ruptures se sont probablement enchaînées, dans des temps plus ou moins long, depuis l'apparition des premières écritures jusqu'aux documents numériques interactifs. En fait le véritable élan de la lecture silencieuse viendrait plutôt des besoins fonctionnels de la croissance de la scolastique au XII^e et $XIII^e$ siècle [135] : la demande intellectuelle de la part des lecteurs pour des longs livres lourdement commentés et parfois contenant un grand nombre de schémas ne pouvait être satisfaite que par la technique relativement rapide de la lecture silencieuse. A partir de là, des procédés spécifiques ont pu se développer avec une relative autonomie par rapport à la langue parlée.

Sous cet éclairage, le défi lancé en essayant de prendre en compte les spécificités de l'écrit lors de la transposition automatique d'un texte à l'oral prend un caractère quelque peu paradoxal : il s'agit de *revenir à la lecture orale d'un texte exploitant des procédés qui ont probablement émergés grâce au fait que la lecture orale n'était justement plus de rigueur!* Cela dit c'est ce paradoxe qui nous donne quelques chances de relever la part de non transposable, de définitivement perdu par l'opération de transposition, et ainsi la possibilité d'une exploitation intéressante pour la mise en place de stratégies multimodales.

1.2.1. Lecture collective et lecture silencieuse : Histoire d'une autonomie

Des **tablettes** de Mésopotamie à celles, tactiles, d'aujourd'hui, l'évolution du support de textes écrits s'est construite dans le temps, au gré de l'apparition de nouveaux matériaux et révolutions technologiques, et par adaptation à différents besoins et objectifs de lecture. Ce processus de transformation quasi « darwinien » du contenant textuel a naturellement engendré à chaque étape celui du contenu.

Longtemps la lecture ne se concevait que comme essentiellement collective et à haute voix. Dès le II^e siècle avant notre ère, le **volumen** permet d'inscrire un texte sur des livres-rouleaux en papyrus aptes à être déroulés pour en révéler le contenu au fur et à mesure. Bien que cette manière de lire puisse s'apparenter à la lecture de longues pages horizontales sur un écran numérique à l'aide de moyens matériels ou logiciels adaptés, elle ne permet pas d'accéder facilement à un point précis, de revenir en arrière, de comparer des passages distants ou les rouleaux entre eux. L'intérêt de l'écriture réside essentiellement dans sa capacité à coucher la parole à destination d'une communauté d'orateurs éduqués et spécialisés ; aucune nécessité non plus de développer des stratégies de lecture rapide lorsqu'il s'agit de préparer un discours, un prêche ou une déclamation publique. L'expertise et l'objectif aidant, nul besoin de faciliter visuellement la « prise en bouche » de la prononciation, ou encore de la structure énonciative et rhétorique du texte : les symboles se succèdent sans séparations et l'utilisation de marques spécifiques assimilables à de la ponctuation apparaît comme extrêmement rudimentaire.

Dans l'objectif de protéger un secret de fabrication et ainsi renforcer la domination d'Alexandrie et sa célèbre bibliothèque, c'est l'interdiction d'exporter la matière première constituant les rouleaux qui a conduit au déclin du papyrus. Dans un contexte de montée de la religion chrétienne entraî-

nant une forte demande de fabrication de livres, d'autres procédés ont dû être développés par les concurrents, à partir du II^e siècle, grâce à l'exploitation de cuirs d'origine animale. Ce nouveau support, le **parchemin**, a rapidement montré des propriétés très avantageuses : matière première commune, maniabilité, exploitation des deux faces, découpes en parties assemblables. Cette rupture importante a précipité la disparition du Volumen au profit du **codex**, ancêtre du livre papier tel que nous le connaissons ; d'un point de vue fonctionnel cela se traduit par une capacité d'avoir une vue plus regroupée et globale d'un ou plusieurs textes et de faire des recherches plus rapides de passages spécifiques. Ces qualités vont être exploitées à travers un lent développement de procédés graphiques conférant **une autonomie de plus en plus affirmée de l'écrit sur son oralisation** : apparition de l'espace entre les mots au VII^e siècle ; marques de paragraphe (pied-de-mouche) au XI^e siècle ; titres de chapitres, notes référencées en marge et table des matières au XV^e siècle.

Un deuxième point de rupture accentue encore cette évolution avec l'avènement de l'imprimerie et la fabrication de livres à l'aide de caractères mobiles ; nous passerons rapidement sur cette histoire beaucoup plus contemporaine qui a permis la mise en place de processus rapides et automatisés de composition et d'uniformisation des textes. Cette capacité accrue de production à des coûts plus faibles a conduit à la démocratisation du livre, de la sphère publique vers la sphère privée, de grands formats collectifs vers de plus petits, de la taille d'une poche. Cela dit, un analphabétisme largement majoritaire laisse longtemps la place à des lectures collectives et orales fréquentes ; il faudra attendre le XX^e siècle pour observer l'apparition d'une grande variété de formats papiers accessibles au grand public (**journaux**, **magazines**). Cette prépondérance, nouvelle, de la lecture silencieuse, fut propice au développement de nouvelles grilles de compositions visuelles mixant textes et images dans des structures complexes et variées. Une conséquence importante est une modification de la nature même de la notion de lecture. L'auteur, l'imprimeur, l'éditeur et le lecteur, tous peuvent jouer avec cette nouvelle dimension pour casser l'approche linéaire habituelle de son intervention. Dans ces nouvelles formes de « livres », multi-thématiques, multi-rédactionnels et multi-usages, les schémas de lecture sont de moins en moins imposés, tout au plus suggérés. En somme, **la richesse des mises en page rajoute à l'autonomie de l'écrit sur l'oral, celle des lecteurs sur les rédacteurs**.

Une dernière rupture nous viens de la révolution numérique et le développement du Web. A la non-linéarité des parcours de lecture, s'ajoute trois caractéristiques. La première concerne la richesse de présentations multimodales et multimédia (texte, vidéo, sons…), la deuxième est liée aux parcours de lecture interactifs, avec la notion d'hypertexte, et la troisième ajoute une dimension inédite aux documents en les rendant potentiellement dynamiques, temporellement évolutifs. Cette évolution questionne une fois de plus la relation entre la lecture et le texte, puisque ce dernier n'est plus une unité stable aux contours relativement bien définis, mais une construction du lecteur/acteur à partir d'un « empilement » de sources hétérogènes, navigables et optionnelles. **L'interactivité, à son tour, rajoute une forme d'autonomie de la lecture elle-même sur le texte**.

1.2.2. Documents visuellement structurés et linguistique

Parmi les différentes disciplines scientifiques, la linguistique pourrait être une des plus à même d'avoir à charge de distinguer et de comparer les modes de communication écrit et parlé. La distinction de deux supports de transmission d'une information en tant que critère de différenciation de deux objets d'études relativement autonomes n'a pourtant que peu été prise en compte. De manière générale, les facteurs graphiques ont reçu une attention limitée dans le champ de recherche circonscrit par la linguistique moderne. Cette vue provient directement de Saussure, usuellement considéré comme le père fondateur de la linguistique moderne et de la sémiologie, qui a placé l'écriture hors du domaine de la linguistique. Pour lui, l'unique raison d'être de l'écriture est de représenter la parole ; le mot écrit n'est que l'image, la photographie du mot parlé. Sa prise en compte pourrait rendre flou et indirecte l'étude du vrai visage de la langue [46]. Bien que depuis des temps récents un certains nombre de problèmes de terminologie montrent que les facteurs graphiques sont à la périphérie du champ de la linguistique, au moins deux limites anciennes sont encore parfois avancées :

- la primauté de la parole : l'écriture n'est pas un langage, mais seulement une manière d'enregistrer le langage par l'intermédiaire de marques visibles [19].
- la restriction à la phrase : la préoccupation pour la parole s'accompagne de la restriction au niveau de la phrase pour la quête linguistique. Étant donné que la phrase semble être le plus haut niveau auquel les concepts de grammaticalité sont intuitivement admis par les utilisateurs du langage, les études « convenables » des linguistes se restreignaient à la phrase. La construction d'unités plus grandes, comme les paragraphes, est plus un genre de choix rhétorique que l'application de règles grammaticales.

Si la linguistique est une discipline qui a la vocation de prendre en charge la description des phénomènes langagiers, elle peine à tenir compte des évolutions engendrées par la nature du support. Qu'il s'agisse de la langue parlée ou écrite, elle a longtemps négligé d'intégrer aux modèles certaines dimensions d'analyse sous prétexte qu'elles relevaient de manifestations dites de surface et donc périphériques ; le fait que la linéarité de la parole peut être relativisée par l'influence de la prosodie a été finalement accepté pleinement dans le champs circonscrit par les recherches en linguistiques. Par contre, le texte écrit n'est toujours pas suffisamment exploré par cette discipline dans sa dimension « planaire » qui permet une organisation non linéaire, voire, à l'instar de la prosodie, en intégrant l'épaisseur supplémentaire produite par les effets de contraste de la mise en forme. L'étude des textes à l'écran pourrait également prendre en compte la potentialité de sa dimension temporelle. Nous renvoyons à [102] pour une étude détaillée et historique conduisant à cette analyse. Malgré tout, certaines études se sont données comme objectif d'étudier le sens véhiculé par la mise en forme du texte. En ce sens, elles prônent l'existence d'une véritable sémantique morpho-dispositionnelle qu'il parait inévitable de confronter aux recherches impliquant le traitement automatique des textes et des langues ; en particulier lorsqu'il s'agit de transposer automatiquement ces phénomènes graphiques de l'écrit vers l'oral.

2. De l'écrit vers oral : sémantique morpho-dispositionnelle et accès non visuel aux documents numériques

La technologie privilégiée pour lire automatiquement les documents numériques à voix haute est la synthèse de parole à partir de texte. Un système de synthèse de parole à partir de texte enchaîne de nombreuses procédures de calcul dont la première consiste à rendre le texte prononçable en le transformant en phrases organisées en mots. Les concepteurs de ces systèmes ont rapidement pris conscience des difficultés de cette étape lorsqu'ils se trouvèrent confrontés à des sources de texte réelles, variées, voire entachées de fautes orthographiques ou syntaxiques : il est apparu qu'il était nécessaire de traiter un certain nombre de phénomènes tels que les sigles, les abréviations, les acronymes, les divers types de nombres et de chiffres ou encore les symboles non alphanumériques. Face à ces difficultés, afin d'obtenir une chaîne orthographique qui permette une transcription de graphèmes en phonèmes, diverses opérations spécifiques sont le plus souvent appliqués aux textes à travers un module de pré-traitement. Ce n'est qu'à partir de ce « nouveau texte » que sont envisagées la phase de phonétisation, puis celle de synthèse acoustique de la suite de phonèmes. Enfin l'obtention d'une synthèse plus acceptable en terme de proximité à la voix humaine nécessite le calcul des paramètres accentuels et intonatifs adéquats (rythme, pauses, mélodie, intensité).

Les premiers travaux qui se sont consacrés à l'étude et à l'amélioration des traitements linguistiques des textes en vue de leur transposition à l'oral recommandaient déjà un traitement particulier pour prendre en compte certains aspects dits métatextuels (ponctuation, inattendus, sigles...) [39]. Par exemple, le système de synthèse de parole reformulera « Au V^e siècle av. J.C., 1kg de plastique coûtait 30576$. » en « Au CINQUIEME siècle AVANT JESUS CHRIST, UN KILOGRAMME de plastique coûtait TRENTE MILLE CINQ CENT SOIXANTE SEIZE DOLLARS. ». Les plus évolués permettent la reconnaissance de motifs de différents types (abréviations communes, prénoms, titres honorifiques, sigles, dates, heures, durées, numéros de téléphone, monnaies, nombres, chiffres romains, URL, adresses électroniques, fractions, unités de mesure, expressions mathématiques, extensions de fichiers informatiques, émojis) et essaie d'intégrer de manière générale certains effets typographiques dans le processus de mise en son [79]. La synthèse de parole à partir de textes a ainsi connu une évolution remarquable et continue au cours des 50 dernières années [153]. En terme de qualité de la voix de synthèse, l'intelligibilité et le naturel ont été améliorés au point que ces deux critères de jugement deviennent de plus en plus obsolètes. Remarquons que nous n'avons pas trouvé de travaux nous permettant de poser un constat définitif prenant en compte la relation entre accessibilité non visuelle des documents, sémantique morpho-dispositionnelle et nouvelles techniques de synthèse de la parole basée sur

l'apprentissage profond. Cependant, déjà les questionnements de [45], puis les conclusions de mes travaux de thèse et les expérimentations plus récentes de [66] confirment que la lecture non visuelle par synthèse de parole peut rester fortement altérée dès que les documents sont fortement structurés, en particulier en raison de la limitation imposée par le choix de l'unité de traitement : la phrase.

2.1. Unité des traitements textuels

La nécessité du choix de l'unité de traitement du texte s'observe lorsqu'il s'agit d'automatiser certains processus informatiques sur la langue écrite. Prenons pour exemple la synthèse de la parole à partir de texte dont l'unité de traitement classique est la phrase. Dès que le document revêt une mise en page complexe, le caractère essentiellement graphique de la notion de phrase rend peu aisée la transposition des connaissances linguistiques vers une unité sonore aux contours bien différents. La sémantique du message véhiculée par le texte original peut s'en trouver dégradée, voire perdue.

2.1.1. Sémantique linéaire

Une première difficulté consiste à prendre en compte la ponctuation, dont les rôles peuvent combiner de nombreuses formes inextricables. Respect de règles typographiques, de choix stylistiques de l'auteur ou de la respiration narrative, volonté de démarcation ou de distinction d'un segment de texte, sont autant de manières d'exploiter la multitude de signes à notre disposition sans sortir de la linéarité de la ligne de texte. Même lorsque la ponctuation n'intervient qu'aux frontières des segments textuels qu'elle encadre, elle peut porter une dimension pragmatique, a l'instar de certaines marques prosodiques, grâce au sens qu'elle véhicule. Comment, sinon, comprendre l'histoire drôle et un peu crue ci-après (et je prie le lecteur de m'en excuser par avance), qui met en relation par la ponctuation de tels effets, parlés et écrits ?

> *Un mécréant, extrêmement grossier, ayant récemment expiré se voit indiquer par Saint-Pierre qu'il ne pourra profiter du Paradis qu'après un dernier test. Il est alors renvoyé sur terre pour une journée, avec comme obligation de ne pas prononcer plus de trois expressions vulgaires. Alors que l'homme trébuche malencontreusement, il s'exclame :*
>
> *— Merde. Oh ! Merde ! Et puis merde…*

Finalement, comme l'énonçait déjà [9] « on doit considérer que l'écrit n'est pas l'image de la parole mais l'image que la langue écrite, en tant que système autonome, donne de la parole. C'est cette autonomie de l'écriture qui fait qu'il n'existe aucun rapport d'imitation naturelle entre l'énonciation de vive voix et la ponctuation ». Nous considérons également avec [27, 129, 57] que la ponctuation est une matière textuelle qui participe à façonner et articuler le langage écrit et devrait faire partie intégrante de la réalité étudiée par le linguiste.

2.1.2. Sémantique planaire

Le choix du rédacteur peut également se porter sur d'autres solutions en profitant des deux dimensions horizontales et verticales offertes par le support. Le jeu de la disposition des segments textuels dans l'espace de la page, permet une forme d'articulation « par le vide » du langage écrit. La variation des espacements participe à construire de nouvelles unités textuelles de niveau intra- ou extra-phrastique. La question que nous posons est de savoir si ces possibilités relèvent d'une analyse linguistique. Une manière de justifier de notre réponse positive est de raisonner par l'absurde à partir de l'exemple suivant pour analyser la portée de l'attente évoquée par « depuis longtemps » et de l'espérance exprimée par « comme espéré » :

(1) Il n'a pas joué comme espéré, depuis longtemps.
(2) Il n'a pas joué : comme espéré, depuis longtemps.
(3) Il n'a pas joué :
comme espéré,
depuis longtemps.

Encore une fois on observe l'importance de la ponctuation pour construire des nouveaux cadres au discours en détachant des segments [4]. Même sans sortir réellement de la notion de phrase graphique, les 3 exemples précédents expriment pourtant 3 messages probablement différents. (1) fait porter l'attente sur le segment qui précède la virgule et donc sur l'espoir d'une certaine qualité de jeu. (2), par l'introduction des deux-points, à tendance à réduire le segment visé par l'attente, et donc circonscrit celle-ci à l'espoir d'une non participation au jeu. Mais au-delà, on observe que (2) et (3) ne diffèrent qu'en terme de disposition de segments textuels et sont absolument identiques quant aux symboles imprimés qui les composent. Pourtant, uniquement par sa forme graphique, (3) permet de faire naître une troisième interprétation potentielle en faisant porter sur la non participation au jeu à la fois l'attente et l'espoir.

Si la modification de la disposition peut insuffler une variation sémantique, nous pouvons arguer que la mise en page participe à construire du sens.

2.1.3. Sémantique spatiale

Un autre degré de liberté offert par le langage écrit s'appuie sur l'investigation d'une troisième dimension d'espace. De la même manière qu'un contour prosodique permet de paralléliser de l'information orale en modelant le signal du message articulable, divers procédés permettent de jouer sur « l'épaisseur », c'est à dire sur des effets contrastifs appliqués aux segments textuels. Par exemple, l'italique sur l'histoire drôle précédente (2.1.1) permet de conférer un statut spécial à ce segment de texte extra-phrastique ; tandis que les contrastes visuels du segment de texte constituant le titre de cette sous-section construisent une unité intra-phrastique.

Plus nous considérons de dimensions spatiales propres aux textes écrits, plus nous remarquons l'absence de leur prise en compte dans les différentes approches linguistiques et les solutions d'Interactions Homme-Machine impliquant du Traitement Automatique des Langues.

Même si une certaine « raison graphique » [64] a été évoquée depuis longtemps, une seule tentative [163] est a notre connaissance suffisamment développée pour poser un cadre théorique et opérationnel intéressant afin de compléter les approches et l'efficacité des solutions. Nous en décrirons rapidement les tenants et les aboutissants dans la section 2.2.

2.1.4. Sémantique spatio-temporelle

Nous noterons que depuis l'avènement de l'Internet moderne, il est possible d'aller plus loin en envisageant une $4^{ème}$ dimension pour la structuration de pages dites dynamiques. Le rédacteur peut exploiter la notion de temporalité de manière a ce qu'elle soit subie par le lecteur sous la forme d'animation ou de capacité auto-adaptative des pages ; mais également émailler le document de possibilités d'interaction et ainsi multiplier des chemins et des temps de lecture auto-adaptables difficiles à gérer lors d'un accès non visuel [117].

De quelle manière l'analyse de ces structures ponctuationnelles, typographiques, dispositionnelles, dynamiques et interactives permet-elle de définir un modèle linguistique qui les intègre ? Cette question doit être abordée pour donner à notre travail un cadre théorique et des méthodes de traitement automatique qui s'appuient dessus. Les recherches les plus abouties, dans ce domaine peu abordé d'une sémantique morpho-dispositionnelle des textes, s'expriment à travers les concepts définis par le Modèle d'Architecture Textuelle.

2.2. Modèle d'Architecture Textuelle

L'architecture de texte est une composante abstraite du texte et prend source dans la notion de Mise en Forme Matérielle (abréviation usuelle : MFM – [163]) et dans l'hypothèse de l'existence d'une équivalence fonctionnelle entre des phénomènes typographiques, dispositionnels et lexico-syntaxiques. La MFM est un sous-ensemble de propriétés morpho-dispositionnelles du texte, propriétés possédant des équivalents langagiers. Ainsi, les constituants des textes, les objets textuels (définition, énumérations, parties, titres, etc.), sont perceptibles par le jeu de contrastes de la MFM et l'architecture de texte est l'ensemble des objets textuels et les propriétés qu'ils entretiennent entre eux.

L'approche générale de l'architecture textuelle est sous-tendue par les travaux de Z. Harris sur le métalangage [68]. Harris pose que la langue peut être décrite par elle-même, c'est à dire que l'on peut exhiber des phrases à portée métatextuelle explicitant le fonctionnement de la langue et obéissant aux mêmes lois de constructions grammaticales que les phrases de la langue. L'ensemble de ces phrases forme le métalangage. Par exemple, la phrase « Max est sujet de Max mange une pomme » est une phrase du métalangage. Harris pose qu'en règle générale ce métalangage est réduit mais que cette réduction laisse des traces dans la langue.

Ce principe a été appliqué aux propriétés morpho-dispositionnelles du texte : ces propriétés peuvent être décrites à l'aide de phrases métatextuelles (par exemple : « je segmente mon texte en trois parties », « je commence par une introduction consacrée à la linguistique », etc.) et la réduction de ces phrases laisse des traces dans le texte qui sont des propriétés syntaxiques,

typographiques et dispositionnelles. L'ensemble des phrases décrivant les propriétés morpho-dispositionnelles et lexico-syntaxiques forment le métalangage architectural [123].

La formalisation du modèle d'architecture textuelle nécessite de définir plus précisément ces notions, susceptibles d'être réutilisées dans la suite du tapuscrit. Les définitions et notations suivantes sont extraites de [123] et [91].

Objet Textuel (OT) *segment de texte caractéristique, accompagné de son entête s'il en a un (liste, énumération, chapitre, section, paragraphe, introduction, préface, théorème, avertissement, rubrique, etc). Dans un texte, l'objet textuel le plus grand est le texte lui-même.*

Mise en Forme Matérielle (MFM) *il s'agit de l'ensemble des propriétés de réalisation appliquées aux objets textuels. Ces propriétés sont de nature syntaxique (nominalisation, numéralisation, formes interrogatives, etc.), typographiques (caractères, polices, corps, styles, couleurs, etc.) et dispositionnelles (justification, colonnages, marges horizontales et verticales, sauts de lignes, de pages, etc.).*

Architecture *composante abstraite du texte, qui est rendue perceptible par un jeu de contrastes de la mise en forme matérielle : mise en relief, mise en parallèle, etc.*

Métaphrase *phrase du sous-langage spécialisé relatif à l'architecturation textuelle. Elle correspond à la forme discursive d'un phénomène de mise en forme matérielle, exprimant l'intention architecturante sous-jacente à ce phénomène. Une métaphrase réalise un acte de discours à vocation textuelle.*

Unité Textuelle (UT) *segment de texte ne comportant aucun objet textuel, c'est-à-dire un segment entièrement discursif du texte.*

Métadiscours *Suite d'instances de métaphrases qui entretiennent entre elles des relations de cohérence et de cohésion. Un métadiscours associé à un texte représente l'architecture de ce texte.*

Cette approche a pour conséquence l'existence d'un continuum entre des formes de Mise en Forme Matérielle (MFM) entièrement discursives et des formes graphiques plus compactes : on passe d'une extrémité à l'autre de ce continuum en effaçant (au profit de marques de MFM) ou en reconstruisant (par interprétation) le métalangage architectural.

Nous avons éprouvé le Modèle d'Artchitecture Textuele (MAT) à travers l'étude de l'oralisation des documents fortement structurés. Nous avons ensuite expérimenté nos résultat dans le cadre applicatif de l'amélioration de l'accessibilité numérique pour les personnes non-voyantes.

2.3. Modèle d'Oralisation

C'est dans le cadre de ma thèse consacrée à la transposition automatique à l'oral des structures visuelles des documents numériques que nous avons étudié certaines limites des lecteurs d'écrans utilisés par les non-voyants ; en particulier celles imputables à la technologie de synthèse de la parole à partir de textes embarquée dans ces logiciels spécialisés.

Aujourd'hui encore, les lecteurs d'écrans peinent à intégrer réellement la composante architecturale du texte lors de l'oralisation. Pour ne parler que des marques ponctuationnelles, celles-ci (lorsqu'elles sont traitées) sont en relation quasi-bijective avec les effets que leur traitement produit : une marque entraîne toujours le même effet. Or, par exemple, un simple tiret ne devrait pas aboutir à une oralisation équivalente selon qu'il est utilisé :

— pour marquer le début d'un item d'énumération ;

— comme une marque d'incise – une parenthétique par exemple ;

— en lieu et place d'un séparateur de type « , » - « ; » - . . .

De manière générale, comme intégré par le Modèle d'Architecture Textuelle, une marque ne peut être considérée isolément. Elle est un élément à part entière du texte, pouvant être lié à d'autres éléments par la constitution intentionnelle et signifiante d'une configuration particulière. Nous avons essayé d'attaquer ce problème par ce qui nous semble être les deux principales limites évoquées dans les sections précédentes, appliquées aux technologies de synthèse de la parole à partir de textes : la faible prise en compte de l'autonomisation de plus en plus marquée des textes produits en entrée du système et le choix discutable de la phrase comme unité de traitement de base. La section suivante s'attache à présenter le modèle issu de cette réflexion.

2.3.1. Schéma synoptique

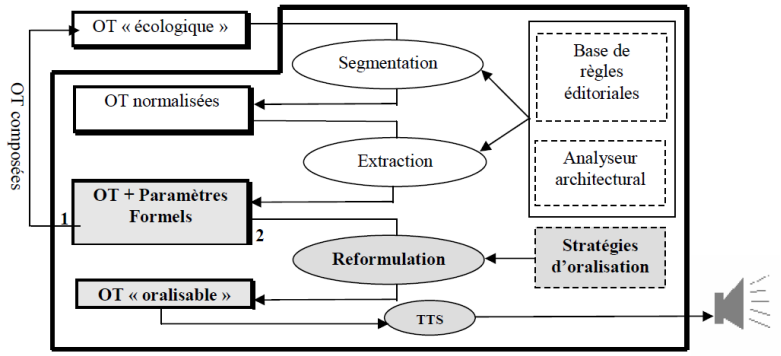


FIGURE 2.1. – Schéma synoptique d'oralisation des structures visuelles des textes

Notre modèle s'appuie en premier lieu sur le Modèle d'Architecture Textuelle et exploite les notions d'Objet Textuel et leur version discursive (Métaphrase) afin de fournir aux lecteurs d'écran une entrée exploitable **sans modification du système de synthèse de la parole intégré**. Le modèle proposé a vocation à être intégré dans une architecture plus large dont un schéma synoptique est décrit par la figure 2.1.

Il s'agit de faire traverser à un texte cible une série de transformations récursives (étapes 1) afin d'en extraire un découpage structuré en OTs normalisés et auxquels sont associés les paramètres formels nécessaires à leur reformulation en phrases analysables par un TTS.

La phase suivante (étape2) exploite un modèle de stratégies de reformulation pour construire un ensemble de métaphrases. Ces dernières pourront être soient directement oralisées par le TTS, soit préalablement transformées en respectant les contraintes et les possibilités de la modalité orale.

2.3.2. Modèle MORTELS

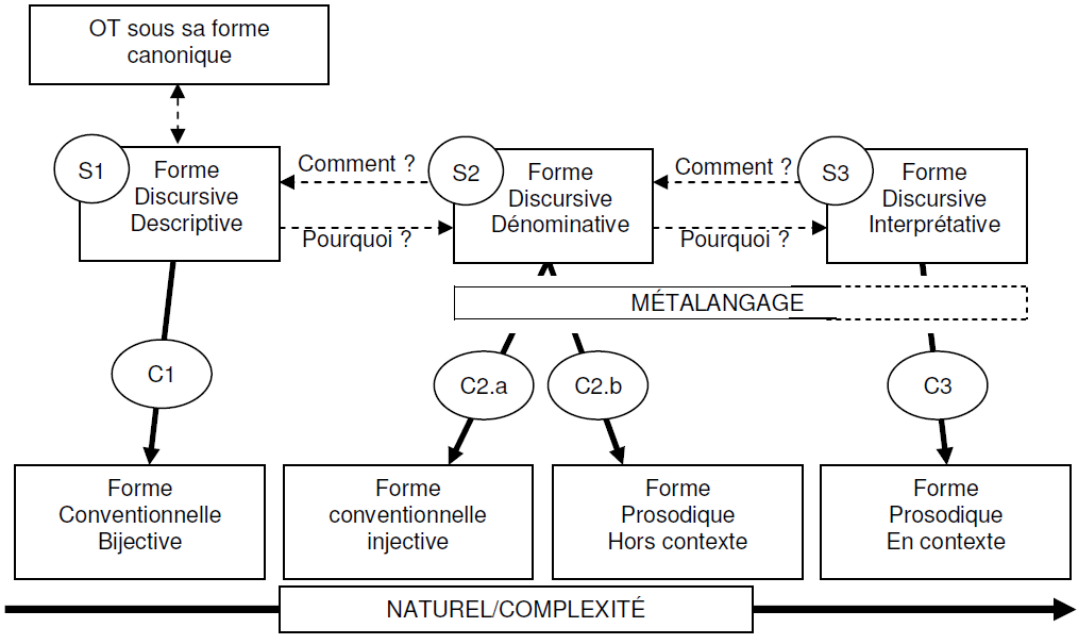


FIGURE 2.2. – Modèle d'Oralisation par Reformulation des Textes Écrits pour être Lus Silencieusement (MORTELS)

L'axiome principal qui sous-tend le modèle prétend qu'il est nécessaire d'obtenir une forme discursive de l'OT commune aux deux modalités et que, de là, pourra être décrit un ensemble de stratégies utilisant des procédés spécifiques à la modalité orale.

Sur ces bases, nous avons conçu le Modèle d'Oralisation par Reformulation des Textes Écrits pour être Lus Silencieusement (MORTELS – figure 2.2).

Le Modèle d'Architecture Textuelle fait l'hypothèse d'une équivalence fonctionnelle entre la Mise en Forme Matérielle (MFM) et un ensemble d'unités textuelles n'intégrant plus que des marques de nature lexico-syntaxiques ; de la même manière, nous stipulons une équivalence fonctionnelle entre ces dernières et des configurations sonores.

Ainsi, la forme discursive peut être considérée comme le pivot central pour l'élaboration d'un continuum de nouvelles formes orales. Celles-ci peuvent être générées par réduction de tout ou partie des marques de MFM de nature lexico-syntaxique, au profit d'un matériel oral non verbal équivalent du point de vue fonctionnel.

La forme discursive d'un OT particulier peut être reconstruite selon trois stratégies :

— par une description objective de la MFM typo-dispositionnelle : forme discursive descriptive (S1) ;

— par un premier niveau d'interprétation, en attribuant un nom à la configuration visuelle : forme discursive dénominative (S2) ; ou

— par une interprétation plus profonde des intentions du rédacteur qui se traduisent par l'utilisation de cet OT particulier : forme discursive interprétative (S3).

Ces trois opérations pourront conduire à 4 continuums de nouvelles formes d'oralisation :

— la forme discursive descriptive pourra se réduire grâce à l'utilisation de conventions « typophoniques » par un procédé d'association bijective entre un ensemble de marques et un ensemble de sons (C1) ;

— la forme discursive dénominative pourra se réduire par une convention « typophonique » qui associe de manière injective les marques définissant un type d'objet particulier à un son unique (C2a) ou par un schéma prosodique spécifique, repéré comme pouvant être associé au type d'OT étudié indépendamment de tout contexte d'énonciation (C2b) ;

— la forme discursive interprétative pourra se réduire par l'utilisation de schémas prosodiques dépendant de la situation d'énonciation et donc de la structure informationnelle de cette forme (en terme de thème, topique, focus, rhème - C3).

Une illustration des stratégies discursives peut être donnée à travers l'exemple du titre :

Méthode prévisionnelle de mise en page d'un ouvrage

S1 : **Le texte en gros, centré et en gras est** méthode prévisionnelle de mise en page d'un ouvrage.

S2 : **Le titre de niveau 2 est** méthode prévisionnelle de mise en page d'un ouvrage.

S3 : **L'auteur va parler maintenant de** méthode prévisionnelle de mise en page d'un ouvrage.

Les parties en gras sont les marques lexico-syntaxiques restituées à partir d'interprétations de complexités croissantes de la morpho-disposition du texte. Ce sont aussi celles qui pourront potentiellement être effacées dans un processus inverse au profit de marques sonores adaptées.

La stratégie S3 permet d'obtenir la forme discursive interprétative, la plus naturelle mais aussi la plus complexe à automatiser. Elle demande une analyse à la fois large et fine de chaque OT afin de construire des typologies indispensables pour expliquer et modéliser la grande variabilité des configurations visuelles possibles. C'est pour cette raison que nous nous sommes concentrés essentiellement sur trois OTs : les mises en saillance intra-phrastiques, les titres et les structures énumératives.

Pour chacun de ces OTs, nous avons pu exploiter une typologie de référence, en déduire les métaphrases constituant leur Forme Discursive Interprétative (stratégie S3) et des règles de réduction au profit de d'une Forme Prosodique En Contexte (continuum de stratégies C3).

Cette dernière opération de transformation s'appuie principalement sur le modèle de l'intonation pragmatique du français développé par Mario Rossi[132] dont nous avons projeté la notion d'intonèmes sur les OTs étudiés. Les configurations prosodiques proposées ont été intégrées dans une synthèse de la parole commerciale. Nous renvoyons à l'article [105] pour une présentation de ce travail.

La section suivante présente les principaux résultats des évaluations perceptives et cognitives par des non-voyants des différentes formes construites. Il m'apparaît important de m'attarder sur ces protocoles expérimentaux car c'est dans leurs conclusions que mes projets de recherche actuels ont trouvées leurs principales sources.

3. Évaluation des premiers modèles et premières limites

Depuis les premières expérimentations de [16] sur les propriétés expressives de la typographie, jusqu'aux travaux plus récents sur l'évaluation de stratégies automatiques [96] ou non [89] de mise en son de structures textuelles, un grand nombre de recherches ont tenté de mettre en regard les effets de signalisation dans les textes, les stratégies de lecture et la cognition humaine. Un état de l'art traversant une partie de cette histoire scientifique peut être trouvé dans [84] et [102].

Le modèle MORTELS et les stratégies qu'il propose avait pour ambition de répondre à des questions du type : *une petite prosodie vaut-elle un long discours ?* Nous avons répondu par l'affirmative dans le champs restreint à quelques Objets Textuels spécifiques (structures énumératives, titres, mise en saillance d'expressions), à condition d'être capable d'en extraire automatiquement les paramètres formels pertinents à leur transformation, d'abord discursive, puis prosodique. Ces nouvelles productions (méta)langagières, basées sur une description de la mise en forme plus ou moins interprétée et plus ou moins discursive, peuvent être considérées comme équivalentes d'un point de vue informationnel. Dans notre quête d'obtenir une véritable équivalence intermodale, la moitié du chemin était parcourue. Une deuxième avancée était nécessaire : étudier l'équivalence de ces stimuli en terme de traitement cognitif (mémorisation / compréhension). Autrement dit, *une petite prosodie vaut-elle* ***mieux*** *qu'un long discours ?*

3.1. Hypothèses et variables

Le protocole visait à évaluer l'efficacité de deux formes interprétatives d'oralisation (discursive - FDI - et prosodique - FPC) en terme d'impact sur le traitement cognitif (mémorisation et compréhension). De plus une forme contrôle (FC) était construite pour chaque stimuli ; elle consistait à une version obtenue par les mêmes effacements dans le matériel textuel pour transformer la FDI en FPC, mais sans contreparties prosodique :

La fabrication de ces trois stimuli a été réalisée de la manière suivante :

— la FDI a été directement oralisée par une synthèse de parole classique ;

— la forme contrôle (FC) a été créée à partir de la FDI en supprimant les segments du signal acoustique correspondant aux marques lexico-syntaxiques « architecturantes » ; ces marques étaient donc effacées et non remplacées ;

— la FPC a repris la forme contrôle mais augmentée de marques prosodiques « équivalentes »

fonctionnellement, selon le modèle MORTELS, aux marques lexico-syntaxiques supprimées.

Deux hypothèses générales étaient testées :

— **H1 :** la restitution de la signification des aspects visuels structuraux de l'écrit améliore les performances en termes de compréhension / mémorisation ;

— **H2 :** la stratégie la plus efficace peut dépendre du type de population envisagé.

De plus, trois variables indépendantes étaient étudiées, potentiellement avec des hypothèses spécifiques associées.

3.1.1. Forme d'oralisation du texte

Deux hypothèses spécifiques relatives à cette variable étaient proposées.

Hypothèse 1 (choix d'une hypothèse de supériorité) : les formes d'oralisation restituant les aspects structuraux produisent de meilleures performances que la forme contrôle qui ne les restitue pas ; comparaison entre la forme contrôle et les 2 autres formes.

Hypothèse 2 (choix d'une hypothèse de différence simple) : les performances des sujets diffèrent selon la forme d'oralisation des textes qu'ils ont écoutée ; comparaison de toutes les formes deux à deux.

Nous n'avons pas proposé de prédiction quant à la supériorité d'une forme sur l'autre car des raisonnements différents permettaient de conduire aux deux hypothèses. Quel élément de décision prévaut pour départager les deux formes entre :

— l'avantage de l'explicitation de l'architecture du texte dans la forme discursive par rapport à la version prosodique plus implicite de ce point de vue ;

— l'inconvénient de la longueur du message pour la version discursive par rapport à la version prosodique qui est sensiblement plus courte ?

3.1.2. Âge des sujets

L'accès aux documents peut être compliqué par des facteurs nécessitant une adaptation des stratégies d'oralisation en fonction des caractéristiques spécifiques à une population d'usagers ; c'est, nous semble-t-il, a priori le cas pour les personnes âgées en raison de potentielles difficultés de perception de la prosodie et/ou de déficiences mnésiques. Ainsi nous proposions de tester à terme les deux hypothèses spécifiques suivantes :

Hypothèse 1 (hypothèse de supériorité) : les performances des sujets jeunes sont meilleures que les performances des sujets âgés ;

Hypothèse 2 (hypothèse d'interaction simple) : l'effet de la forme d'oralisation sur les performances des sujets dépend de leur âge.

3.1.3. Tâche demandée

La mémorisation / compréhension des informations était évaluée à travers des tâches de rappel libre et indicé :

— La tâche de rappel libre consistait pour le sujet à rappeler le maximum d'informations du texte entendu ;

— La tâche de rappel indicé consistait à remplir un texte à trous reproduisant le texte entendu.

3.2. Résultats et tendances

Ces premières expérimentations ont indiqué quelques tendances :

— d'un certain point de vue la forme discursive semble être meilleure que les deux autres formes pour une tâche de mémorisation. En effet, c'est celle où on observe la plus grande proportion de mots justes rappelés, et le moins de mots absents. En notant tout de même que, pour ce qui est de la proportion de mots faux, elle est deux fois plus élevée que pour la forme prosodique mais deux fois moins élevée qu'avec la forme contrôle ;

— D'un autre point de vue la forme prosodique peut être considérée comme meilleure que les deux autres formes pour une tâche de mémorisation, puisque les éléments importants sont retenus bien qu'un grand nombre d'éléments ne soient pas rappelés.

Nous retrouvons, pour conclure vis à vis de ces deux angles de vue, la résistance que nous évoquions pour prédire la supériorité d'une forme par rapport à l'autre (résistance qui avait conduit à opter pour une hypothèse de différence simple). En fait, cette difficulté de prédiction est peut-être corrélée à la nécessité de choisir ce qu'il faut entendre par « une forme est supérieure à une autre » :

— le caractère implicite de la forme prosodique pour évoquer la structure visuelle du texte est accompagné d'une plus grande capacité de marquage que les marques lexico-syntaxiques de la forme discursive ;

— la longueur du message impliqué par l'utilisation d'une forme discursive est accompagnée par une plus grande quantité d'éléments à rappeler que la forme prosodique.

Cela dit, les résultats vont dans le sens d'une validation de l'hypothèse 1 spécifique à la forme d'oralisation : **les formes d'oralisation restituant les aspects structuraux produisent de meilleures performances que la forme contrôle qui ne les restitue pas.**

3.3. Conclusions

La performance des sujets, bien que meilleure si la structure visuelle était prise en compte pour l'oralisation, restait très dégradée par rapport à une lecture visuelle classique. Nous avions démontré l'intérêt applicatif de notre problématique tout en exposant les limites de nos solutions.

Quant à l'objectif scientifique, il s'agissait de franchir les derniers obstacles nous séparant d'une véritable équivalence intermodale, à la fois informationnelle et cognitive. La résistance pour obtenir une équivalence en terme de capacité de mémorisation / compréhension des éléments importants d'un texte a orienté toute la suite de mon travail de recherche.

Si le MORTELS permet de reformuler la structure visuelle, la transposition du texte dans la modalité orale subit les contraintes inhérentes à cette modalité et il semble difficile, dès lors, de restituer les fonctions cognitives étroitement liées à l'exploitation de la spatialité du document et qui agirait globalement au premier regard sur le texte. Bien que la caractéristique anticipatrice de certaines fonctions peut être approchée en restituant très tôt dans le signal de parole certaines marques, l'effet « premier regard » semble bien difficile à reproduire à l'oral.

Ainsi ce travail, à partir de l'étude de la transmodalité de l'écrit vers l'oral, a mis au jour une faiblesse à laquelle nous avons essayé de répondre par une stratégie multimodale. Pour restituer à l'oral la fonction globale d'allègement de la charge cognitive portée par la structure visuelle, ainsi que ses fonctions d'anticipation du traitement du texte, nous avons proposé de distribuer sur deux canaux sensoriels différents le contenu « articulable » et la structure visuelle : les technologies classiques utilisées par les lecteurs d'écran pourraient se charger du contenu de la page tandis qu'une représentation dite par image de page permettrait l'accès à son architecture.

Cette représentation est, à l'origine, un outil de communication et de discussion interne aux concepteurs du Modèle d'Architecture Textuelle, né de la nécessité d'appréhender les architectures plus intuitivement que par les formalismes existants et de favoriser ainsi, à la fois la conception de nouvelles expérimentations et le dialogue lors de collaborations interdisciplinaires. Certains principes ont alors été posés et ont permis d'ériger un système notationnel dédié à la représentation de l'architecture du texte.

Les principes qui définissent une image de page s'appuient sur quatre impératifs auxquels ils doivent répondre : permettre l'observation de phénomènes architecturaux visuels, syntaxiques et rhétoriques ; jouer sur une représentation économique du texte pour ne laisser filtrer que les informations jugées utiles dans un objectif donné ; faciliter l'étude des propriétés des modèles issus de différentes disciplines et la composition de ces modèles ; servir d'outil d'expérimentation. Nous avons imaginé une cinquième direction pour l'utilisation de ce système : la réalisation d'interfaces interactives multimodales pour la communication homme/machine.

La fin de mon travail de thèse a consisté à démontrer la plausibilité cognitive de cette dichotomie page / image de page dans le cadre d'une présentation multimodale du texte. Les expérimentations menées et les résultats obtenus sont résumés dans le chapitre suivant.

4. Dichotomie page / Image de Page et accès non-visuel interactif aux textes

Notre objectif était de permettre l'accès aux documents numériques en dissociant la perception du contenu articulable de la structure purement visuelle. Aussi chaque dimension peut être appréhendée par deux canaux sensorimoteurs différents. Notre première idée fut de développer la notion d'Image De Page Interactive qui pourrait se décliner à terme dans une version tactile, permettant un accès non visuel à l'architecture des textes.

La notion d'Image de Page a été initié par [92] et développé par [72]. Ce concept consiste à une opérationnalisation d'un langage notationnel de l'architecture textuelle. Les auteurs décrivent 5 types d'Image De Pages selon la finesse que l'on veut donner à la granularité de la représentation :

— l'image minimale qui représente l'organisation du texte en unités textuelles (UT) en conservant la casse du début des UT (M/m pour respectivement majuscule/minuscule) et la ponctuation encadrante. La suite de caractères constituant les UT est signifiée par un trait horizontal ;

— l'image augmentée 1 qui ajoute à la précédente les blancs utilisés pour rendre perceptible l'architecture de l'objet textuel représenté (indentation, marge, interlignage, ...). Ces marques sont représentées par des flèches verticales et horizontales de hauteur et longueur relative ;

— l'image augmentée 2 qui ajoute à la précédente les segments de texte mis en saillance dans l'OT représenté (guillemets, gras, italique, ...) et les marques lexico-syntaxiques (organisateurs textuels, connecteurs, ...) ;

— les images augmentées 3 et 4 qui s'attaquent respectivement à la représentation des relations de dépendance syntaxique et à la représentation des relations de dépendance rhétorique.

La granularité des deux dernières représentations semblait trop fine dans le cadre de cette étude puisqu'elles relèvent d'un niveau de description de la structure informationnelle de l'OT qui va au delà de ce que le premier regard est a priori susceptible de nous fournir. Aussi, nous pensons que l'image augmentée 2 permettait une représentation de l'OT intéressante pour un système de présentation de l'information. La première nécessité fut de s'assurer de l'utilisabilité de ce principe. Nous avons mené deux études expérimentales visant à évaluer si l'Image De Page constitue une représentation plausible sur le plan cognitif, c'est-à-dire manipulable mentalement par le sujet.

4.1. Premier protocole

L'étude que nous avons menée visait à évaluer si le sujet humain comprend la représentation du texte sous la forme d'une IDP en testant la faisabilité d'une tâche impliquant sa manipulation. Nous avons choisi d'évaluer cette compréhension en demandant au sujet de choisir, parmi plusieurs IDP, celle qui représente le mieux un texte oralisé par un système de synthèse vocale à partir de textes.

4.1.1. Hypothèses et *design*

Notre hypothèse générale était que le sujet humain est capable d'associer à un texte oralisé sa représentation sous la forme d'une IDP. Pour tester cette hypothèse générale, la méthode que nous avons employée a consisté à faire écouter à chaque sujet deux formes orales d'une énumération, exploitant la stratégie discursive interprétative (FDI) et la forme prosodique en contexte (FPC). Simultanément à l'oralisation du texte, trois IDP étaient présentées au sujet sur l'écran de l'ordinateur. La tâche du sujet était de classer les IDP par ordre décroissant en fonction de leur capacité à représenter le texte oralisé.

Plus spécifiquement, voici quelles étaient nos hypothèses :

— **H1 :** les sujets sont capables d'identifier une énumération à l'oral, qu'elle soit réalisée sous une forme discursive ou prosodique et ils ont des connaissances sur la réalisation des énumérations à l'écrit. Ils sont alors capables d'associer une réalisation adéquate d'une énumération sous la forme d'une IDP à une énumération entendue à l'oral ;

— **H2 :** les sujets sont capables de juger une représentation sous la forme d'une IDP incompatible si celle-ci représente un autre objet textuel que l'énumération ;

— **H3 :** les sujets sont capables de repérer qu'une énumération oralisée sous une forme discursive peut être représentée par deux IDP :

  ◦ sous la forme d'une image de page représentant une suite de phrases. Il s'agit alors d'une réalisation non redondante puisque les marques structurales sont réalisées dans la structure énumérative sous une forme lexico-syntaxique ;

  ◦ sous la forme d'une image de page représentant une énumération marquée par un pattern morpho-dispositionnel « lassique ».

Trois IDP étaient présentées au sujet :

— une IDP dite « Structure Énumérative » (SE) exploitant un pattern morphodispositionnel « classique » (figure 4.1 à gauche) ;

— une IDP dite « Bloc » n'exploitant aucune marque de mise en forme morphodispositionnelle spécifique à un objet textuel particulier. Cette IDP est possible pour représenter l'énumération, mais elle n'en constitue pas une représentation spécifique (figure 4.1 au centre) ;

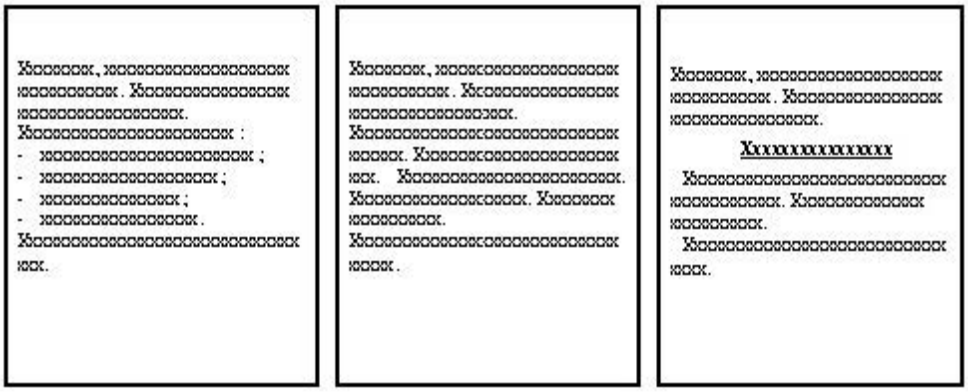


FIGURE 4.1. – 3 Images De Pages (IDP) à classer par pertinence selon le stimulus écouté (FDI ou FPC) ; De gauche à droite, IDP SE, Bloc et Autre.

— une IDP dite « Autre » présentant des marque morpho-dispositionnelles pouvant indiquer un titre ou une structure d'insistance dans le texte. Cette IDP est supposée ne pas représenter l'énumération du fait qu'elle n'en constitue pas une représentation adéquate (figure 4.1 à droite).

Trois prédictions découlent des trois hypothèses formulées plus haut :

— **P1 :** quelle que soit la forme oralisée entendue, l'ordre de classement favorise « l'IDP SE » en premier choix ;

— **P2 :** quelle que soit la forme oralisée entendue, « l'IDP Autre » est plus choisie en troisième choix, qu'en deuxième et qu'en premier choix ;

— **P3 :** la forme oralisée discursive (FDI) favorise davantage le choix de « l'IDP Bloc » en premier que la forme prosodique (FPC).

4.1.2. Résultats

Quelle que soit la forme entendue, les sujets favorisent un premier choix portant sur « l'IDP SE » dans leur classement. L'étude de leurs justifications montre que tous les sujets ayant fait ce choix le justifient par une correspondance entre l'énumération identifiée à l'oral et celle représentée à l'écrit et ceci de façon explicite. Ainsi, les sujets sont capables d'identifier l'objet textuel « énumération » à l'oral, d'identifier ce même objet textuel représenté par une IDP, et d'associer les deux.

« L'IDP Autre » est davantage choisie en troisième qu'en deuxième ou en premier, quelle que soit la forme entendue. Les sujets sont capables d'identifier l'énumération à l'oral et de constater l'incompatibilité avec l'image représentant un titre ou une insistance. Les catégories de justifications avancées pour le troisième choix portant sur « l'IDP Autre » renforcent cette explication ; en effet, sur 26 troisièmes choix portant sur « l'IDP Autre », 21 justifications évoquent une mauvaise correspondance entre le message entendu et ce qui est représenté sur l'image de page. Cinq

sujets évoquent un choix par défaut.

L'écoute de la forme discursive de l'énumération favorise davantage en premier choix « l'IDP Bloc » que la forme prosodique. Ce résultat n'est vrai que lorsque la forme prosodique est écoutée en premier. Deux hypothèses explicatives peuvent rendre compte de cette interaction entre la forme écoutée et sa position dans l'ordre de présentation des deux formes. Une première explication est que lorsque l'ordre de présentation est FPC-FDI, lors de l'écoute de la FPC, le sujet a n'a eu aucune difficulté à y associer l'image de page SE ; lors de sa deuxième écoute portant sur la FDI, il aurait « envie » de donner le même ordre de classement avec « l'IDP SE » en premier choix mais il se sent obligé de changer d'ordre. Si ce changement d'ordre se fait au hasard, il devrait y avoir alors autant de premier choix portant sur « l'IDP Bloc » que sur « l'IDP Autre » associée à l'écoute de la forme discursive. Or, ce n'est pas le cas : sur 7 sujets ayant associé à la forme prosodique écoutée en premier, un premier choix de « l'IDP SE », 6 sujets ont choisi « l'IDP Bloc » pour la FD. Il semble donc que le changement d'ordre conditionné par l'ordre d'écoute des formes ne se fasse pas au hasard mais en faveur de « l'IDP bloc ».

Une explication alternative est que l'effet d'interaction est dû à la différence de statut prise par la forme prosodique selon sa position dans l'ordre de présentation ; lorsque le sujet écoute en premier la forme discursive, ce qu'il perçoit en priorité, c'est la présence d'une énumération compatible avec la représentation de l'image de page SE ; il choisit donc en premier cette image. Lorsque le sujet écoute en premier la forme prosodique, il associe sans difficulté l'image de page SE à la forme prosodique ; lorsqu'il entend ensuite la forme discursive, il perçoit que l'énumération est différente de celle entendue sous une forme prosodique. Prenant celle-ci pour référent, il choisit une autre image que celle choisie en premier pour la forme prosodique et son choix porte sur celle qui est possible bien que moins probable, « l'IDP Bloc ». L'étude des justifications données par les sujets semble confirmer cette interprétation de l'interaction ; en effet, les 6 sujets ayant choisi en premier « l'IDP Bloc » considèrent que cette image correspond à la forme discursive car elle représente une suite de phrases. Certains sujets comparent la forme discursive à la forme prosodique qu'ils ont écoutée avant et évoquent le fait que « c'est des phrases plutôt que des points ». En revanche, un résultat obtenu semble contradictoire avec nos hypothèses : la forme discursive, lorsqu'elle est écoutée en premier, donne uniquement lieu à des premiers choix portant sur « l'IDP SE », et donne autant de seconds choix portant sur « l'IDP Bloc que Autre », alors que dans la logique de nos hypothèses, elle aurait dû davantage donner lieu à des deuxièmes choix portant sur « l'IDP Bloc ». L'étude des catégories de justifications évoquées pour un deuxième choix portant sur les « IDP Bloc ou Autre » pour la forme discursive écoutée en premier n'éclaire pas ce phénomène.

Malgré ce dernier résultat peu compatible avec nos hypothèses, l'hypothèse générale selon laquelle le sujet humain est capable d'associer au texte oralisé sa représentation sous la forme d'une IDP semble validée. Les résultats de ce premier protocole nous permettent donc d'envisager une présentation multimodale du texte combinant l'oralisation du texte et sa représentation visuelle sous la forme d'une image de page.

4.2. Second protocole

La prise d'information visuelle globale et quasi-instantanée (dite de *skimming*) précède la mise en place de stratégies efficaces de consultation ou de recherche rapide plus locale dans un document (dites de *scanning*) . En particulier, ces stratégies pourront s'appuyer sur la capacité du premier regard à permettre d'anticiper les diverses compositions selon ses objectifs de lecture et son expertise du document.

4.2.1. Hypothèses et prédictions

Le protocole visait à la fois à renforcer nos résultats sur la plausibilité cognitive des IDP et à valider la valeur « désambiguïsatrice » de nos stratégies : il s'agissait pour le sujet de choisir parmi deux IDP celle qui correspond le mieux à l'oralisation du texte écouté selon 3 formes d'oralisation : Les formes discursive (FDI), posodique (FPC), et contrôle (FC) conçues de la même manière que pour les expérimentations décrites dans le chapitre précédent.

Une oralisation partielle de ces 3 stimuli est présentée aux sujets (l'amorce et le premier item d'une énumération). La question était de savoir si le sujet est capable de déclencher lui-même à la fin de l'écoute le raisonnement selon lequel il s'agit de l'oralisation d'un début de texte (un entraînement est effectué sur un texte complet et il est simplement précisé qu'il s'agira ensuite du même type d'exercice).

Afin de renforcer l'ambiguïté de notre stimulus et susciter ainsi un effet négatif sur le traitement syntactico-sémantique de la structure du message (effet dit de *garden path*) nous avons construit un texte qui conserve une cohérence sémantique malgré les amputations subies (voir figure 4.2). Les IDP associés à l'écoute de chacune de ces formes sont celles de la figure 4.3.

Les **trois** points forts de la région Midi-pyrénées sont (énumérés ci-après.) (Le premier point fort comprend) les **trois** cultures agricoles locales.

FIGURE 4.2. – Exemple de source permettant de créer les 3 stimuli FDI, FPC et FC ; en rouge les éléments amputés pour la FPC et la FC

FIGURE 4.3. – IDP associées aux stimuli oraux

La prédiction est que la FDI sera supérieure ou égale à FPC ; elle même très supérieure à la forme contrôle :

— malgré la reprise du numéral dans le premier item qui semble appuyer l'interprétation d'une phrase autonome, la forme prosodique permettra de conférer au message, dans les mêmes proportions que la forme discursive, le statut de début d'énumération ;

— malgré un texte d'entraînement sur un texte complet et sans précision supplémentaire, la plausibilité cognitive des images de page est assez forte pour permettre d'inférer le statut d'un objet textuel partiellement représenté par ce moyen.

4.2.2. Résultats

Les résultats vont dans le sens de nos prédictions :

— les sujets appartenant aux groupes des formes discursives et prosodiques ont tous su « entendre » le début d'une énumération et inférer le statut de début d'IDP correspondante ;

— les sujets appartenant au groupe de la forme contrôle ont à 80% déduit de l'écoute une IDP correspondant à une phrase autonome.

Cette étude nous a ouvert des perspectives pour l'utilisation des IDPs dans un contexte d'interaction homme/machine. Plus généralement, l'idée d'exploiter la structure visuelle en la dissociant perceptivement du contenu articulable a pu naître de notre démarche scientifique initiale :

1. étude de la transposition d'une modalité vers une autre en terme de pertes/gains à la fois d'un point de vue informationnel et cognitif ;
2. compensation des pertes informationnelles en développant des modèles de génération ;
3. compensation des pertes cognitives par des stratégies multimodales adaptées.

Si le point (1) avait comme cadre applicatif naturel l'accès aux documents numériques par les non-voyants ; le point (3) a pu nourrir des idées d'applications pour tous. En particulier pour pallier la difficulté de l'accès à l'information sur très petits écrans en combinant des Images De Pages Interactives représentant les pages Web et l'oralisation des segments des IDPs pointées par l'utilisateur.

C'est avec la perspective d'exploiter ces nouveaux cadres que mes recherches ont migré vers Caen et le laboratoire GREYC. En particulier, commençait à naître l'idée de profiter des avantages évoqués dans les deux points précédents : projeter la notion d'Image De Page vers un canal sensoriel non visuel pour faire profiter les non-voyants d'une multimodalité dissociant la structure visuelle des documents et le contenu articulable produit par les lecteurs d'écran.

La deuxième partie de ce document s'attache à présenter les étapes de concrétisation de cette approche dans mon nouvel environnement de travail et de collaborations.

Deuxième partie

Stratégies de lecture non visuelle : rapide, globale, interactive

Contexte et environnement :

- *Laboratoire de recherche en sciences du numérique de Caen (GREYC)*
- *Équipes Interaction, Sémiotique : LANgue, Diagrammes (ISLAND - 2005 à 2006) puis Documents Langues Usages (DLU - 2007 à 2011) puis HUman Language TECHnologies (HULTECH - 2012 à 2021)*
- *Maître de conférence - publications 2007-2020 - 2 thèses en co-encadrement*

Les expérimentations avec les non-voyants durant mon travail de thèse ont mis en évidence les limites d'un modèle basé sur la reformulation pour conserver à la fois toute la sémantique véhiculée par la mise en forme mais également les capacités de compréhension et de mémorisation qu'elle participe à faciliter. Une des perspectives était de considérer un même document selon deux partis pris distincts : celui d'un contenu articulable logico-thématiquement structuré mais également celui de l'image d'une page visuellement contrastée. Cette dernière qualité participerait fortement à l'activation d'une capacité de vision globale et au développement de stratégies de lecture de haut niveau. Une fois cette dichotomie page/image de page acceptée, si le premier parti pris reste compatible avec des stratégies orales de reformulation, l'autre peut s'appuyer sur un accès non visuel au travers de la modalité tactile. C'est de cette idée qu'est né mon premier projet soutenu par l'Agence Nationale de la Recherche (ANR) en 2013, en tant que coordinateur scientifique : Accès par Retour Tactilo-oral Aux Documents Numériques (ART-ADN).

Les récentes conclusions de ce projet mettent également en évidence, dans les conditions expérimentales que nous avons pu mener, un certain nombre de limites. L'accès tactile à la mise en forme permettrait de développer des stratégies intéressantes pour la recherche rapide d'éléments (*scanning*) dans un document mais insuffisantes pour développer celles qui s'appuieraient sur une vision globale et rapide (*skimming*). Après deux Contrats Plan Etat Région (CPER) d'évaluation de sa faisabilité, nous développons actuellement un nouveau projet, TAGTHUNDER, qui a trouvé un financement par le Fonds national pour la Société Numérique (FSN), géré par la Banque Publique d'Investissement (BPI) dans le cadre de l'appel « Accessibilité numérique » du Plan Investissement Avenir 2 (PIA2). Il consiste à faire retrouver une telle aptitude non visuelle de « premier regard » en construisant des versions sonores spatialisées de pages Web intégrées dans des systèmes interactifs appropriés.

Je détaillerai le cheminement scientifique qui a abouti à cette idée à travers les points suivants :

- Herméneutique, énaction et interprétation des textes : vers l'étude de systèmes de substitution sensorielle automatique ;
- Application au projet TactiNET : métaphore de la canne blanche ;
- Application au projet TagThunder : métaphore de la *cocktail party* ;
- Résultats autour de la segmentation automatique de pages Web.

5. Herméneutique, énaction et interprétation des textes : vers l'étude de systèmes de substitution sensorielle automatique

L'insertion dans mon nouvel environnement de recherche Caennais au sein de l'équipe Island du laboratoire GREYC, n'a pu véritablement s'exprimer dans ces premières années qu'à travers la constitution d'un groupe de travail normand, intitulé Groupe ν (*i. e.* **NU** pour **N**ouveaux **U**sages). J'ai ainsi intégré un consortium d'enseignants-chercheurs des universités de Caen et Rouen, travaillant dans les domaines des sciences pour l'ingénieur, de l'informatique, de la linguistique et des sciences cognitives. Dans une perspective d'aide à l'interprétation des textes et avec une démarche centrée utilisateur, il s'agissait de considérer la constitution du sens comme activité sémiotique centrale dans les interactions homme-machine [53]. En particulier, observant l'essor constant d'environnements numériques de travail enrichis par de nouvelles interfaces de visualisation ainsi que par des technologies sophistiquées pour la recherche et l'indexation d'information, nous cherchions à croiser des approches en TAL et en reconnaissance de formes dans des IHM centrées sur l'interprétation de l'utilisateur. L'objet était de questionner la nature d'interactions menées par l'usager dans un environnement favorisant l'émergence de son interprétation (plutôt que dirigées voire imposées automatiquement par le dispositif informatique) [52]. Cette ouverture à la créativité et à la sérendipité des interfaces reconnaît une interaction toujours singulière et donc difficile, sinon impossible, à modéliser ; cette difficulté interroge encore aujourd'hui les chercheurs à propos de la place de l'utilisateur et de son interprétation dans les couplages personne/système et personne/environnement.

J'ai découvert dans l'herméneutique en tant que science de l'interprétation des textes [44] le courant d'une herméneutique dite **matérielle** qui résonne de manière intéressante avec les approches dont j'ai développé les tenants dans la première partie de ce document ; ne serait-ce que par les analogies terminologiques avec le Modèle d'Architecture Textuelle qui permet l'analyse et l'interprétation des propriétés syntaxiques, typographiques et dispositionnelles dites de mise en forme elle aussi **matérielle**. Dans les deux cas, cette adjectivation en matérialité se justifie notamment parce qu'elle engage une réflexion sur l'unité des deux plans du langage : la relation entre le contenu d'une part et son expression d'autre part, à travers l'analyse des traces produites par les différents acteurs du texte (rédacteur, éditeur, lecteur, interacteur...) [14]. Cette considération globale de l'environnement et des conditions de production et d'accessibilité des textes m'a naturellement conduit a m'intéresser avec le groupe ν à la démarche énactive (proposée par le biologiste Franscico Varela pour penser la cognition à partir des organismes

vivants) [101]. Posant les conditions d'efficacité d'une boucle perception/action vertueuse, la théorie de l'énaction m'a permis de tisser les liens conceptuels entre tous les aspects de ma problématique autour des interfaces langagières de lecture rapide non visuelle. En particulier, j'ai pu trouver une source d'inspiration dans certains travaux remarquables issus de cette approche lorsqu'ils sont appliqués au développement de systèmes de substitution sensorielle [170, 60, 56].

Dans ce cadre, ma participation aux réflexions épistémologiques du groupe ν autour d'une science des textes instrumentée a posé les premières briques constitutives des **cinq piliers qui soutiennent ma recherche actuelle**. Je n'ai eu de cesse, durant les dix années qui ont suivi, d'enrichir cette construction scientifique pour en consolider au mieux les fondations et la cohérence.

5.1. De la pluridisciplinarité

Ma sensibilité à une démarche pluridisciplinaire, riche et ouverte, a perduré jusqu'à aujourd'hui, depuis le programme cognitique du CNRS dans lequel évoluait mon travail de thèse et ma double qualification dans les sections 07 et 27 du Conseil National des Universités. Cependant, j'observais déjà à cette époque le paradoxe entre d'une part les incitations du CNRS pour le développement de programmes et de nouvelles sections spécifiques au recrutement de profils intrinsèquement pluridisciplinaires, et d'autre part la difficulté d'intégration de ces profils dans des équipes de recherche parfois touchées par le syndrome soupçonneux du « touche à tout, bon à rien ».

Dans mon domaine de spécialité actuel et avec l'avènement plus ou moins récent des techniques innovantes basées sur l'apprentissage machine et en particulier l'apprentissage profond, le traitement automatique des langues connaît également une évolution tournée vers une pluridisciplinarité excluant de plus en plus les sciences humaines. Mes récentes expériences de conférences internationales de premier rang m'indiquent qu'à l'intérieur même du champs circonscrit par la notion très large de sciences du numérique, il est parfois difficile de rassembler plusieurs communautés ; peu voire pas d'interfaces langagières à VRST2019 (*Virtual Reality Software and Technology Symposium*) pour rapprocher les chercheurs en TAL et en IHM ; plus étonnant encore à ICDAR2019 (*International Conference on Document Analysis and Recognition*), pas de session plénière associant spécialistes du texte et du document, ce dernier étant essentiellement couvert par les chercheurs en traitement automatique d'image ; les apports mutuels des deux traitements automatiques, des langues et des images, semblent pourtant particulièrement naturels et prometteurs pour aborder les problématiques d'analyse de documents numériques en tenant compte en même temps des contenus et de leurs propriétés expressives.

Dans le développement de mes propres recherches, j'ai également recentré un temps mes efforts sur des problématiques d'interaction Homme Machine et d'apprentissage automatique en mettant de côté mes accointances linguistiques. Je n'ai pour autant jamais négligé les collaborations avec les sciences cognitives qui ont même été au centre de la relance de mes activités de recherche dans cette période de transition. C'est sous l'impulsion du pôle pluridisciplinaire Modélisation en sciences cognitives (MODESCO) de la Maison de la Recherche en Sciences

Humaines de Caen que j'ai trouvé les ressources collaboratives et financières pour construire un projet autour de la perception tactile, présenté au chapitre suivant, à trois niveaux d'interdisciplinarité :

— niveau partenarial local : (1) laboratoire de Psychologie des Actions Langagières et Motrices (PALM, aujourd'hui Laboratoire de Psychologie de Caen Normandie - LPCN) en tant que partenaire local spécialiste des démarches expérimentales ayant trait à la santé et à l'éducation, (2) le Centre de recherches inter-langues sur la signification en contexte CRISCO qui est à l'origine du développement de la synthèse de la parole à partir de textes KALI longtemps utilisée par une communauté de non-voyants, et (3) l'équipe électronique du GREYC pour sa dimension à la fois technique et experte de la perception des vibrations basse-fréquence ;

— niveau financier régional : intégration au programme NUMNIE (Contrat Plan État Région 2015-2020), dispositif transversal SHS/STIC situé sur un des axes prioritaires de l'université de Caen Normandie. Cet espace de convergence pluridisciplinaire porté par le pôle Document numérique a permis de valoriser nos premier travaux et les pousser vers un soutien national ;

— niveau financier national : responsabilité scientifique du partenaire multiporteur GREYC dans le cadre d'une réponse favorable d'un appel à projet transveral de l'ANR (programme Sociétés innovantes visant à favoriser la coopération et la confrontation des approches des disciplines des sciences humaines et sociales et des problématiques soulevées par les autres disciplines scientifiques).

5.2. De l'autodétermination

Les travaux scientifiques développant des connaissances se voulant utiles à plus ou moins long terme à la prise en charge de handicaps, doivent, plus que d'autres, s'appuyer sur des études de besoin sérieuses et documentées. Un des enseignements que j'ai pu recevoir à la fois au contact des chercheurs en situation de handicap qui composaient ma première équipe d'accueil à l'IRIT, puis des usagers rencontrés et interviewés lors de différentes démarches expérimentales, concerne le besoin d'autonomie par l'autodétermination de ses choix.

Cette réflexion est probablement une fausse évidence puisqu'elle n'est pas souvent présente dans les définitions invoquées par les spécialistes du *design for all*. Ce concept issu des travaux de [61] a conduit à la définition de nombreux systèmes de principes, règles voire normes d'accessibilité. Il est par exemple souvent fait référence encore aujourd'hui aux 7 grands principes régissant les conditions d'accessibilité telles que définies dès 1997 par le centre pour la conception universelle de l'université de Caroline du Sud [1] :

1. **Utilisation égalitaire** par des personnes ayant différentes capacités ;
2. **Flexibilité d'utilisation** répondant à une vaste gamme de préférences et de capacités individuelles ;

1. https://projects.ncsu.edu/ncsu/design/cud/about_ud/udprinciplestext.htm

3. **Utilisation simple et intuitive** permettant une compréhension facile, indépendamment de l'expérience, des connaissances, des compétences linguistiques ou du niveau de concentration de l'utilisateur ;
4. **Information perceptible** pour une communication efficace, quelles que soient les conditions ambiantes ou les capacités sensorielles ;
5. **Tolérance pour l'erreur** afin de réduire au minimum les conséquences des actions involontaires ;
6. **Effort physique minimal** pour une utilisation efficace et confortable, générant une fatigue minimale ;
7. **Dimensions et espace libre pour l'approche et l'utilisation** afin qu'il soit aisé de s'approcher, saisir, manipuler et utiliser quelles que soient la taille, la posture ou la mobilité.

Si ce type d'énumération peut être un guide intéressant pour évaluer a posteriori un outil, une technologie ou une interface, et produire des recommandations et des aides automatiques à la conception, elles pourraient peiner également à générer des solutions de rupture ; ne risquent-elles pas parfois d'enfermer le concepteur, et donc avec lui l'utilisateur, dans des solutions fonctionnelles mais trop restrictives, simples voire simplistes et nivelées par le bas ?

En projetant cette réflexion sur notre problématique de l'accessibilité non visuelle aux pages web, une critique essentielle que nous formulons à l'endroit de nombreuses solutions de la littérature porte sur l'approche méthodologique qui est posée. Elles s'attachent, pour la plupart, à réduire la charge cognitive en recherchant dans la page Web les informations « pertinentes » et en éliminant les perturbations induites par les éléments « périphériques » [2, 20] ; ou encore en intégrant des techniques de résumés[122]. De manière générale il s'agit de simplifier le contenu pour le rendre plus digeste aux modalités tactiles ou orales. Ce faisant, le concepteur considère que l'amélioration de l'accessibilité doit sacrifier une certaine richesse de contenu.

Nous pouvons cependant trouver quelques définitions récentes plus générales mais finalement plus proches de notre état d'esprit vis à vis de la notion d'accessibilité universelle. Elles affirment qu'elle est « le caractère d'un produit, d'un procédé, d'un service, d'un environnement ou de l'information qui, dans un but d'équité et dans une approche inclusive, permet à toute personne de réaliser des activités de façon autonome et d'obtenir des résultats [au moins] équivalents »[2]. C'est cette prise en compte de la capacité d'autodétermination de l'utilisateur d'un système interactif qui nous semble majeure et prioritaire à évaluer. L'ambition est de ne pas s'identifier subjectivement à un usager théorique dont les capacités et les intentions ne sont pas modélisables, mais plutôt concevoir des modèles permettant de fournir à l'utilisateur réel l'ensemble de l'information dans sa complexité ; à la charge du concepteur de trouver les stimuli adéquats pour favoriser l'interaction et le déroulement de la boucle perception/action ; à la charge de chaque utilisateur d'apprendre à les maîtriser, à se les approprier, à pérenniser ses propres interprétations ; à la charge du temps et de la pratique de faire émerger des stratégies nouvelles de prélèvement d'information.

2. Définition développée en 2011 par le Groupe DÉFI Accessibilité dans son Rapport de recherche pour les milieux associatifs de Montréal – Accessibilité universelle et designs contributifs (version 5.3), LANGEVIN, ROCQUE, CHALGHOUMI et GHORAYEB, Université de Montréal, `https://raamm.org/laction-du-raamm/laccessibilite-universelle/`

5.3. Du *design for more*

La définition de la section précédente s'attache à circonscrire le champs d'investigation du *design for all* en y intégrant la notion d'autodétermination de ses choix par tout usager d'un système interactif. Nous pourrions arguer également que nombre de dispositifs ingénieux émaillent la littérature spécialisée dans l'accessibilité mais peu trouvent le chemin du transfert industriel pour réduire concrètement la fracture numérique.

Selon nous, deux cas peuvent se présenter et soutenir l'analyse de cette situation.

Le dispositif peut parfois ouvrir à une accessibilité nouvelle en autorisant des activités totalement empêchées jusque là. La problématique relève alors essentiellement de la viabilité du processus industriel associé. Survient dans ce cas la difficulté du cloisonnement de ces solutions avec un public restreint et donc d'une rentabilité hasardeuse de leur développement. Un exemple frappant souvent cité, issu du monde entrepreneurial, est celui de l'OPTACON [60] qui autorisait dès le milieu du XX^e siècle l'accès au livre papier pour tous les non-voyants ; cela sans faire appel à des codes alternatifs tel que le Braille. Un stylo scanner permettait de suivre ligne par ligne l'encre imprimée d'un texte en produisant dynamiquement sous la pulpe d'un doigt de l'autre main la sensation tactile de la forme des lettres survolées. Après un apprentissage de quelques semaines, un utilisateur pouvait accéder à une lecture d'environ 16 mots à la minute. Malheureusement, regrets partagés par les usagers non-voyants que j'ai pu rencontrer, la production et la maintenance de ce dispositif américain innovant, efficace et utilisé s'est arrêtée faute de rentabilité financière.

Pour limiter ce risque, une approche orientée *design for all*, que nous partageons dans nos travaux, s'appuie sur un décloisonnement des recherches tournées vers les handicaps sensoriels en les rapprochant de problématiques issues de contraintes liées à la situation ou l'environnement d'usage. C'est dans ce sens que nous développons notre projet de CEcité TExtuelle et LEcture Multigrain (CETELEM) qui vise à élargir la notion de non-voyance et de mal-voyance physique à des difficultés de lecture visuelle d'ordre cognitif (dysphasie spécifique, forme d'autisme, complexité informationnelle) ou d'ordre situationnel (pas ou peu d'écran). Pour autant, nos questionnements ne rentrent pas dans le cadre évoqué par le premier point puisque des solutions de lecture non-visuelle existent déjà, en particulier à travers l'utilisation des lecteurs d'écran. Que cela soit sur ordinateurs de bureau (avec JAWS ou NVDA et une multimodalité parole/Braille) ou sur dispositifs tactiles (avec VoiceOver ou Talk Back et la modalité orale), les technologies d'accessibilité des pages Web n'ont cessé de s'améliorer et des stratégies de lecture plus ou moins efficaces peuvent continuer de se développer. Cela dit, l'amélioration des technologies n'implique pas l'amélioration de l'accessibilité elle-même en raison de l'évolution concomitante de l'enrichissement et de la complexité des supports. De notre point de vue, des limites importantes restent saillantes, en particulier pour répondre au besoin d'accès à des stratégies de lecture interactive qui s'appuient sur une perception des documents à la fois non visuelle, globale et rapide.

Dans cet autre cas, pour lequel l'objectif est davantage d'optimiser une technologie existante ou de l'augmenter de nouvelles possibilités, d'autres écueils nous semblent aujourd'hui assez repérables pour pouvoir être évités. En particulier le risque est de remplacer une technologie

comportant limites reconnues et possibilités maîtrisées, par une autre, plus puissante pour un objectif spécifique mais ne prenant pas en charge certaines fonctionnalités utiles de l'ancien dispositif. Il sera dans ce cas difficile de faire accepter l'amélioration proposée si elle est au détriment d'usages bien intégrés. Nous prônons dans nos travaux actuels la conception d'Interactions Homme Machine non destructrices qui viennent compléter une offre existante ; avec la perspective évoquée dans la section précédente, que l'usager saura de lui-même combiner, sélectionner, contourner les fonctionnalités à sa guise voire parfois au grès de résultats fortuits efficaces.

Un moyen de favoriser un peu plus cette appropriation des interfaces est mis en évidence par certaines recherches originales. Celles-ci montrent que la froideur opérationnelle des technologies proposées peut également être un frein à l'acceptabilité de la solution [85]. Trouver des approches qui intègrent une composante émotionnelle et relationnelle est certainement un défi majeur encore peu abordé. Nous travaillons dans ce sens autour de propositions qui prennent un soin particulier à construire des environnements virtuels assez cohérents et sensoriels pour déclencher une forme de plaisir émotionnel.

Finalement, ce qui caractérise notre approche est de favoriser l'acceptabilité des interfaces pour tous et l'interprétabilité des objets interactifs non pas en limitant, en recentrant ou en prédigérant les usages mais au contraire en additionnant à l'existant **plus** de situations, **plus** de fonctionnalités, **plus** d'interactions et **plus** de sensibilité. C'est cette conception que je défendrai dans les projets présentés dans les chapitres suivants sous la dénomination de *design for more*.

5.4. De la sérendipité et de l'émergenge

Parfois la création de connaissance n'est rendue possible que par la rencontre d'un hasard apparent avec la sagacité d'un observateur avisé. Le terme qui recouvre ce phénomène est un néologisme naît d'un voyage terminologique passant par la Perse, l'Italie, la France et l'Angleterre en 1754 pour revenir finalement en France dans les années 1980 sous le vocable de sérendipité, mais recouvrant un champ de signification très ouvert [3]. Seulement en 2014, une définition générale est proposée pour ce mot français avec « l'art de découvrir ou d'inventer en prêtant attention à ce qui surprend et en imaginant une interprétation pertinente ». [147]

De nombreux chercheurs ont essayé de construire une typologie de ce phénomène en croisant les niveaux d'intentionnalité, de chance ou d'inattendu ayant conduit à de très nombreuses découvertes majeures. Ces propositions, recensant selon les auteurs de 2 à 40 types de sérendipité [154, 150, 152], sont presque toujours orientées par le point de vue de l'inventeur dont la recherche de solution est le métier. Pourtant il est fréquent d'observer ce phénomène à l'oeuvre dans le couplage utilisateur / système interactif. Le hasard ou les erreurs de manipulations, combinés à la sagacité interprétative des utilisateurs du système produisent des détournements d'usages qui prennent parfois une portée collective. Une explication de la capacité de généralisation de ces nouveaux usages peut être trouvée dans un modèle de l'émergence. Pour le

3. Le mot *serendipity* a été proposé par l'écrivain anglais Horace Walpole, inspiré du conte oriental « Voyages et aventures des trois princes de Serendip » de l'italien Cristoforo Armeno traduit par le français Louis de Mailly.

concepteur de systèmes interactifs, la question se pose alors de construire des environnements qui favorisent le rôle de la sérendipidé et l'émergence des changements organisationnels. Pour le scientifique il s'agit de proposer un cadre théorique pour l'étude de cette question ; il doit permette la construction d'expériences et de nouvelles connaissances afin d'en évaluer la portée et la pertinence. Le chercheur en informatique essaiera plus précisément de saisir et de modéliser la portée algorithmique d'une automatisation de bout en bout de la démarche ; et dans notre cas, son impact sur les problématiques d'analyse automatique du texte et de l'activité de lecture, objets centraux et indissociables de notre intérêt.

Dans mes travaux visant la transposition de stratégies de lecture visuelle vers des modalités tactiles et/ou orales, les cadres expérimentaux propices à de telles études et sur lesquels nous fondons notre approche s'appuient sur une conception dite énactive de la perception spatiale des formes sonores ou tactiles. Aussi, nous ne souhaitons pas questionner directement les représentations internes, tant est qu'elles existeraient, qu'un utilisateur plongé dans un environnement interactif (le navigateur Web) se ferait de son espace d'action (la navigation) et de perception (les pages Web). En nous plaçant à l'interface entre Traitement Automatique des Langues et Interaction Homme Machine nous interrogerons d'abord, et parfois de manière volontairement radicale, l'activité de couplage sensoriel et moteur entre un utilisateur et un environnement peuplé de contenus interactifs tactiles et/ou sonores. La qualité des stimuli perçus, parcourus et interprétés pendant cette navigation non visuelle est mesurée à l'aune de leur capacité à faire émerger (à énacter) une sémantique interprétative non plus soumise à une représentation interne, générique et contraignante mais consubstantielle à une perception active, située et riche.

Par le choix de nos cadres scientifiques et expérimentaux, nous nous inscrivons dans la continuité des questionnements autour des systèmes dits de substitution sensorielle ; en particulier nous trouvons une force d'inspiration et de conviction dans les résultats publiés dès la fin des années 1960 par Paul Bach y Rita et son fameux *Tactile Vision Substitution System*, ou TVSS [12]. Nous reviendrons dans les chapitres suivants sur les liens de filiation entre nos propositions et cette lignée de travaux, développés à destination des personnes aveugles. Ils ont posé des résultats à la fois fondamentaux et spectaculaires dans la manière d'aborder les relations entre les notions de perception et d'action.

5.5. De la créativité

La première difficulté de notre approche réside dans la construction *a priori* d'environnements sensoriels et de situations interactives qui favorisent la fluidité de la boucle perception/action. Pour donner corps à un tel cadre d'expérimentations et de recherches nous pensons nécessaire de maximiser en premier lieu la cohérence des stimuli perçus, l'affordance des objets interactifs proposés et le naturel de l'analyse par l'usager lui-même des situations rencontrées. Dans cette optique, nous justifions une utilisation quelque peu radicale de l'analogie ou de la métaphore dans l'art d'élaborer de la connaissance en exploitant sa capacité potentielle à structurer le système conceptuel humain [81]. Bien que son rôle dans une perspective autre que la médiation scientifique puisse être discutable, nous pensons que lorsque la force de la tension métaphorique est bien ajustée, dans un contexte de découverte, elle participe légitimement et

efficacement au travail d'inférences d'hypothèses heuristiques.

L'utilisation et le filage des métaphores présentées dans les sections suivantes s'inscrivent dans cette démarche. Leur construction relève d'un processus d'analyse et de compréhension des spécificités de notre domaine d'étude particulier : l'amélioration de l'accès en lecture rapide et non visuelle des textes écrits. Les travaux présentés dans la première partie nous ont rapidement convaincu de l'intérêt d'une perception globale des contenus articulables en même temps que de leur capacité expressive. C'est en affinant ce constat et notre définition de certains processus cognitifs mis en jeu dans la lecture dite silencieuse que se sont forgés les projets en cours et à venir. En particulier, toujours avec l'objectif d'exploiter les propriétés visuelles des documents dans de nouvelles modalités, les métaphores présentées s'appuient sur les notions de *skimming*, de *scanning*, d'alternance fluide entre accès local *vs.* global et de *Gestalt* non visuelle.

Les propriétés visuelles du texte agissent sur de nombreuses dimensions : (1) elles jouent sur la lisibilité des documents et donc sur leur accessibilité cognitive ; (2) à l'instar de la prosodie, elles véhiculent une part de la sémantique du message ; (3) par la qualité d'affordance qu'elles procurent au document, elles exploitent nos tendances perceptives naturelles pour suggérer des parcours interprétatifs cohérents ; (4) elles génèrent des possibilités nouvelles propres aux intentions de lecture individuelles. En cela, elles soutiennent l'émergence de stratégies créatives par expérience ou par sérendipité ; (5) Elles favorisent la capacité d'action de l'œil à combiner rapidement des opérations de prélèvement d'information à la fois locales et globales. C'est cette interaction et cette dynamique, qui sous-tend toutes les autres, que nous souhaitons nous employer à conserver lors de la transposition des propriétés visuelles dans de nouvelles modalités sensorielles.

Nous nous intéressons plus précisemment à deux stratégies procédant d'un enchaînement de processus mentaux fréquents lors de la lecture silencieuse de documents par un voyant : le lecteur jette un premier regard sur tout ou partie de la page et en effectue un survol quasi-instantané (*skimming*) ; il initie ensuite une recherche rapide d'indices visuels et langagiers, sélectionnés en fonction des intentions de lecture (*scanning*) ; ces deux stratégies, plus ou moins conscientisées, peuvent se répéter selon différentes combinaisons jusqu'à la satisfaction d'objectifs individuels. La mise en page et la typographie prennent une part déterminante dans le succès et l'efficacité de ces processus. Notre réflexion porte sur la possibilité de les rendre accessibles aux non-voyants ; autrement dit, comment favoriser le développement de stratégies de *skimming* et de *scanning* non visuels en s'appuyant sur une transposition orale et/ou tactile de la structure visuelle des documents ?

Les sections suivantes décrivent deux projets qui ont l'ambition de participer à répondre à cette question de transposition sensorielle en soutenant notre créativité expérimentale.

Le premier projet, **TactiNET**, est guidé par la métaphore du concept de canne blanche projeté dans l'environnement des dispositifs mobiles et tactiles : le non-voyant explore le monde en se dirigeant grâce aux contacts de sa canne avec les obstacles et les matériaux autour de lui ; nous souhaitons que la sémantique morpho-dispositionnelle et les contrastes lumineux qu'elle induit sur l'écran puisse jouer ce rôle pour l'exploration des documents par transformation en stimuli vibratoires et thermiques. Le paysage tactile résultant, fait de « trottoirs textuels » et de « chemins texturés », alimentera les mouvements de notre « canne-doigt ».

L'analogie qui sous-tend le développement du second projet, **TagThunder**, est un prolongement de la métaphore connue par les psychologues sous le nom de « *cocktail party effect* ». En psycho-acoustique, elle dénote la possibilité de focaliser son attention sur un flux verbal dans l'ambiance bruyante d'une réception ; que cela soit dirigé vers ses interlocuteurs ou vers des sources extérieures à sa conversation. Nous filerons cette métaphore en considérant la relation entre le lecteur non-voyant et les zones de la page Web, comme celle entre un invité situé au centre d'une salle et les différents groupes de discussions qui s'y sont formés : les échanges sont séquentiels à l'intérieur d'un groupe mais concurrents entre les différents groupes ; l'invité, nouveau venu, doit prélever dans ce paysage sonore suffisamment d'informations pour identifier rapidement la discussion à laquelle il souhaite se mêler.

6. Application au projet TactiNET : métaphore de la canne blanche

Le travail décrit dans les sections de ce chapitre a été réalisé sous ma responsabilité scientifique dans le cadre du projet ANR **A**ccès par **R**etour **T**actile **A**ux **D**ocuments **N**umériques (**ART-ADN**) en partenariat avec le laboratoire de **P**sychologie des **A**ctions **L**angagières et **M**otrices (**PALM**). Une publication scientifique récente dans une revue internationale [104] détaille les principaux tenants et aboutissants de cette recherche ; une valorisation du projet et de ses résultats peut être visionnée sur *https://www.youtube.com/watch?v=jAqYQzfL-is&t=9s*.

L'objectif général de cette recherche est de permettre l'accès à la structure visuelle d'une page Web dans un environnement non visuel ; et ainsi augmenter le message avec de l'information véhiculée par une part de la sémantique morpho-dispositionnelle souvent peu accessible.

Le projet **TactiNET** est uns solution guidée par la métaphore du concept de canne blanche projeté dans l'environnement des dispositifs mobiles et tactiles : le non-voyant explore le monde en se dirigeant grâce aux contacts de sa canne avec les obstacles et les matériaux autour de lui ; nous souhaitons que la sémantique morpho-dispositionnelle et les contrastes lumineux qu'elle induit sur l'écran puisse jouer ce rôle pour l'exploration des documents par transformation en stimuli vibratoires et thermiques. Le paysage tactile résultant, fait de « trottoirs textuels » et de « chemins texturés », alimentera les mouvements de notre « canne-doigt ».

Dans ce but, nous avons développé le dispositif vibrotactile **TactiNET**, qui convertit la structure visuelle d'une page web en paysages tactiles pouvant être explorés sur la plupart des dispositifs mobiles. Les contrastes lumineux survolés par les doigts sont capturés dynamiquement, envoyés à un microcontrôleur, traduits en motifs vibratoires variables en intensité, fréquence et température, puis reproduits par nos actionneurs sur la peau à l'endroit défini par l'utilisateur. Ses principales caractéristiques sont décrites par la figure 6.1.

De nombreux dispositifs ont été proposés pour fixer des actionneurs vibrotactiles sur le corps des utilisateurs afin d'augmenter la perception et la mémorisation des informations. L'idée d'un système de substitution sensorielle dynamique peut être trouvée dès les années 1920 comme mentionné dans [33]. Dans le cas spécifique de la transposition d'informations visuelles sous forme de stimuli tactiles, une série d'expériences remarquables sont décrites dans [12], qui a inventé le terme Tactile Vision Sensory Substitution (TVSS) à cette fin. Dans ce cas, 400 actionneurs (solénoïdes) sont intégrés dans le dossier d'une chaise et l'utilisateur assis dessus manipule une caméra pour scanner les objets placés devant lui. Les images capturées sont ensuite traduites

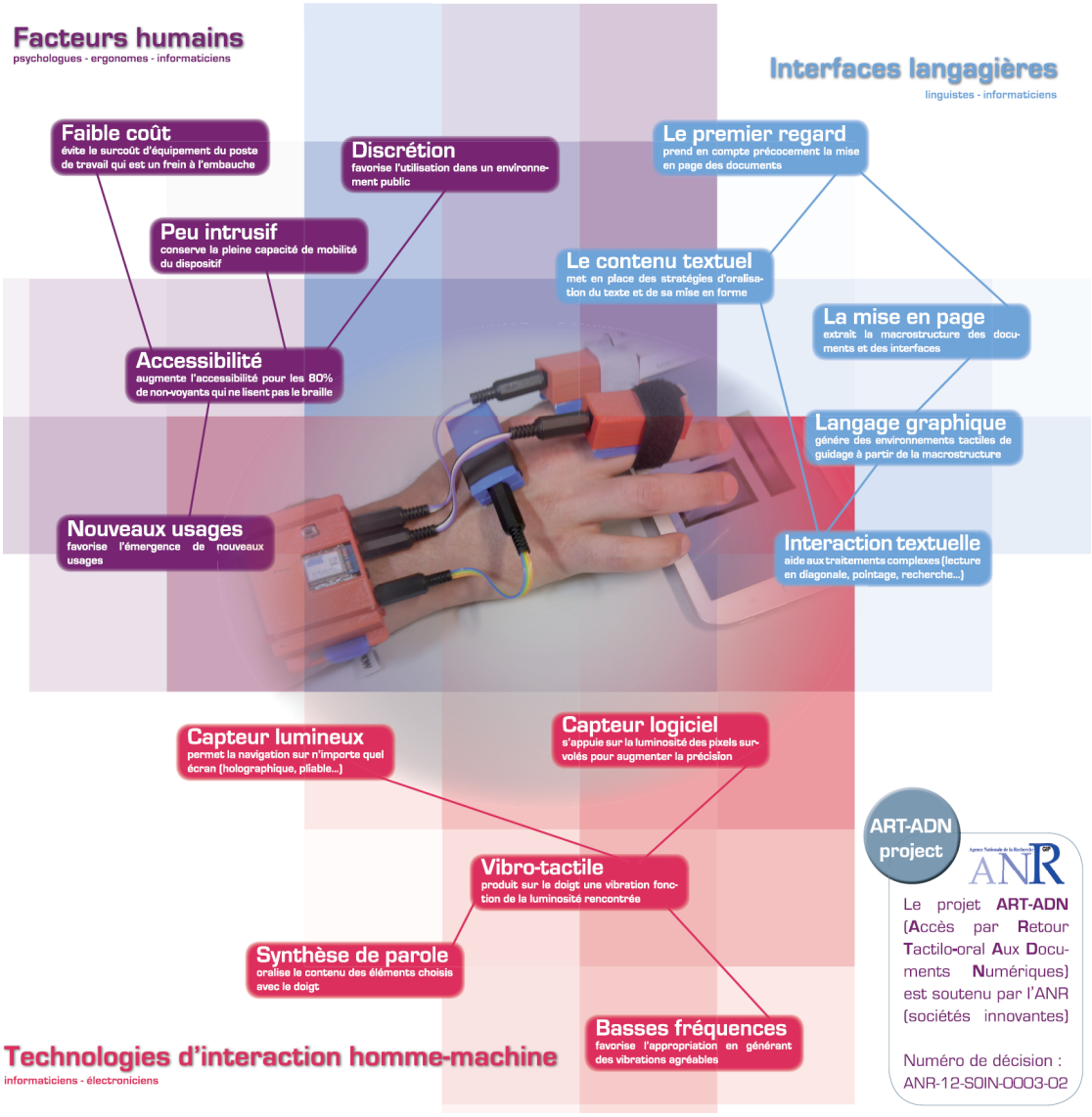


FIGURE 6.1. – Caractéristiques du dispositif TactiNET.

en stimuli vibratoires qui sont transmis dynamiquement aux actionneurs. Les résultats spectaculaires de ces expériences démontrent le pouvoir de la plasticité du cerveau humain pour (1) exploiter naturellement des informations visuelles encodées dans une modalité sensorielle de substitution, (2) extérioriser ses sensations et (3) créer du sens d'une manière comparable à celle qu'aurait produit la perception visuelle. Quelques années plus tard, l'Optacon a été proposé pour offrir un retour vibrotactile [60]. Ce dispositif particulièrement innovant est capable de transposer des textes imprimés en stimuli vibrotactiles. Un stylet optique et une règle permettent de suivre les lignes d'un texte et de reproduire les formes des lettres de manière dynamique sous

la pulpe d'un doigt positionné sur des picots vibrotactiles. Trois semaines d'apprentissage en moyenne suffisent pour obtenir une performance de lecture d'environ 16 mots par minute, ce qui constitue un résultat remarquable car il ouvre la possibilité d'accéder à des livres papier non spécifiques, c'est-à-dire conçus pour les voyants. L'idée intéressante est de soutenir la substitution sensorielle par une exploration active et une transposition analogique plutôt que symbolique (par rapport au braille, par exemple dans le système D.E.L.T.A [37]). Malheureusement, malgré son très bon accueil à l'époque par la population aveugle, la commercialisation du produit a été courte, faute d'un équilibre économique durable. Plus récemment, avec l'avènement des nouvelles technologies portables et l'augmentation constante de la puissance des applications et des actionneurs embarqués, nous pouvons observer de nombreuses études intéressantes pour l'utilisation du toucher dans de nouvelles interactions. [141] révèle l'intérêt de notifications tactiles riches sur les smartphones avec divers emplacements pour les actionneurs, et différentes formes de vibration. [97] présente un dispositif simple et peu coûteux qui intègre la modalité haptique dans l'interaction. Les auteurs se concentrent en particulier sur la manière dont une personne peut interagir avec la friction, la taille, la texture et la malléabilité des objets numériques. D'autres dispositifs de recherche exploitant les idées développées dans le cadre de TVSS sont dédiés à l'amélioration de la perception des personnes aveugles pour faciliter leur déplacement autonome. Dans ce but, [41] propose d'extraire automatiquement les contours des images capturées dans l'environnement de la personne aveugle. Ces informations sont ensuite transposées en stimuli vibratoires par un ensemble de 48 moteurs intégrés dans un gilet porté par le sujet. Ainsi, ce système de navigation (Tyflos) intègre un réseau de vibrations 2D, qui offre à l'utilisateur aveugle une sensation de l'espace environnant en 3D.

Dans une approche orientée vers la tâche, ces propositions ne couvrent pas adéquatement nos besoins concernant la prise en compte de la typographie et de la mise en page pour la lecture non visuelle des pages web. Un certain nombre de recherches plus spécifiques se rapprochent de nos perspectives. Certaines ont porté sur l'utilisation de textures pour produire différentes sensations tactiles lors de l'exploration spatiale de graphiques [70, 71, 100], conduisant à des recommandations sur les paramètres de composition en termes de forme élémentaire, de taille, de densité, d'espacement, de combinaison ou d'orientation. D'autres dispositifs ont été développés pour faciliter l'accès aux diagrammes par les personnes aveugles, en particulier Tactos [155] mais aussi [127, 86, 59, 121]. Cependant, ils n'ont pas vocation à s'attaquer à la complexité globale des documents web multimodaux qui peuvent rassembler des informations textuelles, visuelles, de mise en page, pour n'en citer que quelques-unes. Un premier navigateur web tactile adapté aux documents hypertextes a été proposé par [133]. Des filtres sont appliqués aux images et aux graphiques pour réduire la densité des informations visuelles et extraire les informations importantes à afficher sur un écran tactile dédié. Le texte est rendu sur le même écran tactile en codage braille à 8 points. Ce navigateur illustre les trois principales limites que nous souhaitons lever. Premièrement, il ne prend en compte qu'une partie des informations de mise en page qui ne sont pas suffisantes pour exploiter la richesse des phénomènes typographiques et les contrastes lumineux qu'ils induisent. Deuxièmement, le navigateur utilise le braille pour rendre le texte à l'écran alors que seule une minorité de personnes aveugles peut le lire ; et qu'il est de toute manière limité pour traduire les informations typographiques et de mise en page. Troisièmement, l'autonomie de l'utilisateur est réduite dans le choix des informations accessibles

puisque le navigateur décide unilatéralement des informations susceptibles de l'intéresser. Un autre navigateur a été proposé, CSurf [95], mais il repose également sur le filtrage des données et les informations intéressantes sont sélectionnées par le navigateur lui-même. TactoWeb [128] est un navigateur web multimodal qui permet aux personnes malvoyantes de naviguer dans l'espace des pages web grâce à un retour tactile et audio. Il s'appuie sur une cellule qui stimule le bout du doigt en étirant et contractant la peau latéralement. Bien qu'il préserve les positions et les dimensions des éléments HTML, TactoWeb trie et adapte les informations sur la base de l'analyse de la structure de l'arbre DOM (Document Object Model) de la page Web et non sur son organisation visuelle réelle. Proche de l'idée du système Tactos [155] appliqué aux pages web, le navigateur proposé par [80] requiert une souris tactile pour communiquer la présence de motifs HTML survolant le curseur. La souris est équipée de deux cellules tactiles positionnées sur le dessus. Au cours des évaluations du système, des pages web ont été présentées aux participants, puis explorées à l'aide du dispositif. Il a été demandé à chaque utilisateur aveugle de décrire la mise en page des pages visitées. Les résultats indiquent que si la mise en page globale semblait être perçue, la description révélait encore des incohérences dans la relation entre les éléments et dans la perception de la taille des objets. L'idée d'une perception tactile analogique d'une page web est séduisante mais seulement si le vocabulaire tactile du dispositif est suffisamment riche pour transposer correctement la sémantique visuelle des pages web. De plus, l'inconvénient d'utiliser un navigateur spécifique brise partiellement notre principe de « design for more » sur lequel nous souhaitons baser notre solution. Une des raisons qui fragilise l'appropriation d'un dispositif par une personne aveugle est son aspect destructeur. Nous faisons l'hypothèse qu'un outil, même s'il offre de nouvelles fonctionnalités utiles, risque de ne pas être accepté s'il empêche l'utilisation de fonctionnalités largement éprouvées. Le système doit pouvoir être ajouté et combiné aux outils classiquement utilisés par un individu donné, que ce soit avec une synthèse vocale, une plage braille ou un navigateur spécifique.

Nous noterons que des recherches récentes s'attachent toujours à améliorer l'accessibilité des surfaces tactiles numériques qui « surpeuplent » notre quotidien d'une multitude de services domotiques, téléphoniques ou urbains. Par exemple [119] s'attache à limiter les conséquences de cette dissémination sur la population d'usagers malvoyants ; les technologies actuelles ont en effet une résolution insuffisante en terme de propagation des vibrations sur la surface des écrans et ne permettent pas une perception localisée et adaptée à une expression tactile sémantique riche. Leur approche est de ne pas contourner ce problème en déportant les stimuli sur des accessoires externes (gants, vestes) mais en concevant des méthodes de retour haptique multipoint localisé sur une surface. Leur technologie *LotusBraille* [120], basée sur la méthode d'entrée *Perkins Brailler* permet déjà de transmettre avec de bons résultats des lettres en braille sur les doigts des deux mains en contact avec une surface tactile.

Dans tous les cas, qu'il s'agisse d'améliorer l'accessibilité des dispositifs existants ou de concevoir de nouveaux périphériques d'interaction efficients, un triple objectif doit être envisagé selon nous pour développer des outils performants dans le contexte de l'accès non visuel à l'information : la cohérence des perceptions, la fluidité des actions et la génération d'émotions. Dans notre cadre particulier, nous faisons l'hypothèse que l'amélioration des lecteurs d'écran est intimement liée à la perception de la cohérence de la structure visuelle des pages Web, tant pour les informations contenues, que pour les interactions qu'elle suggère et la valeur émotionnelle

qu'elle véhicule. Nous rajoutons que ce système idéal ne doit pas entraver les mouvements exploratoires, être autonome, robuste et léger. Il devrait également être discret et peu coûteux. Nous avons conçu le dispositif **TactiNET** pour aller dans le sens d'une réponse à ces contraintes.

6.1. Description du dispositif

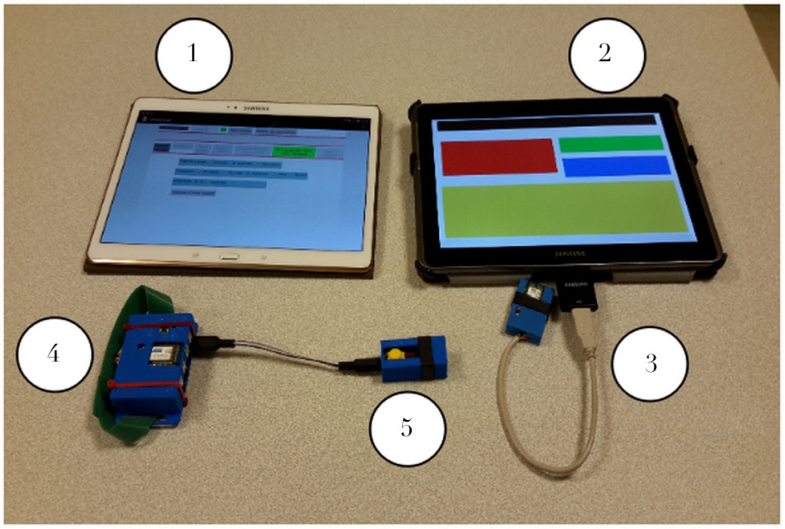


FIGURE 6.2. – éléments du dispositif TactiNET

Conçu et réalisé entièrement au laboratoire GREYC, en collaboration avec l'équipe Électronique, le dispositif TactiNET [104] a été développé pour offrir à la fois polyvalence et facilité de paramétrage lors de la conception d'expériences. Comme le montre la figure 6.2, il comprend :

— une tablette de contrôle (item1) manipulée par l'expérimentateur dans laquelle tous les paramètres tactiles peuvent être gérés et programmés (par exemple la forme des motifs, les fréquences de vibrations, etc.) avant d'être envoyés par Bluetooth à la tablette de l'utilisateur testé (item2) ;
— une tablette utilisateur (item2), dans laquelle les pages Web sont traitées et affichées selon un langage graphique. Lors d'une navigation tactile sur l'écran, les coordonnées des cinq doigts de l'utilisateur sont envoyées au système hôte via une connexion fournie par un dongle ZigBee (item3) ;
— le système hôte (item4) auquel les actionneurs peuvent être connectés. Chaque actionneur peut être soit un vibreur piézoélectrique (item5) pour fournir un retour haptique, soit un dispositif Peltier pour fournir un retour thermique.

En fonction des informations survolées par les doigts, les informations sont transmises de la tablette de l'utilisateur au système hôte, qui peut alors commander les actionneurs piézoélectriques et le Peltier en fonction des valeurs d'amplitude, de fréquence et de température

demandées. Le système hôte est alimenté par batterie avec un système de gestion USB intégré. Il est donc portable et offre plus de 2 heures d'expérience avec une charge complète.

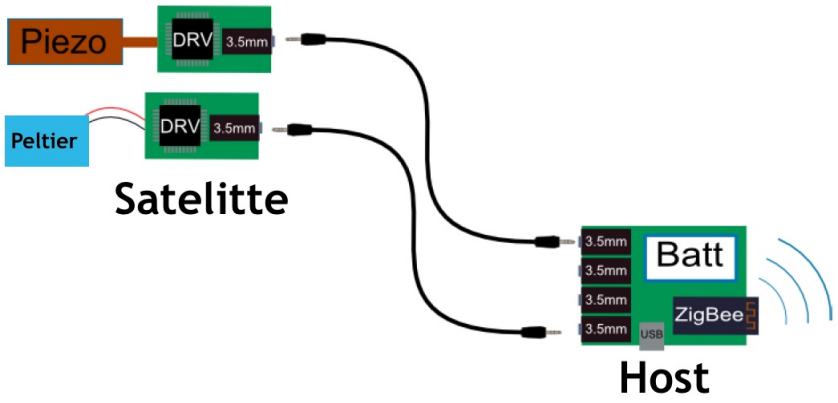


FIGURE 6.3. – Le système hôte communique avec la tablette de l'utilisateur via ZigBee. Les satellites sont connectés via des connecteurs jack 3,5 mm et contiennent des actionneurs piézoélectriques ou Peltier.

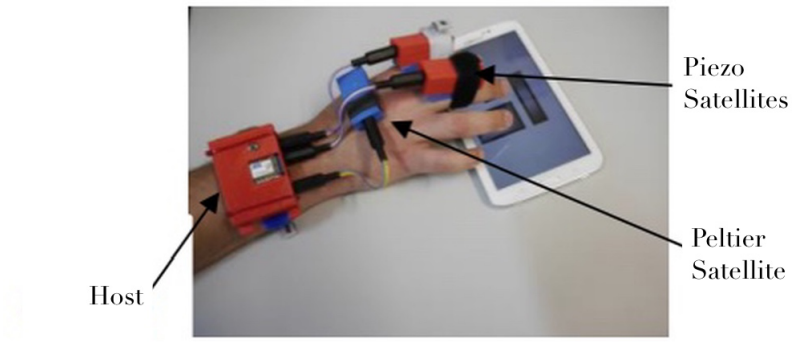


FIGURE 6.4. – Exemple de montage sur main active avec une base, deux piézoélectriques et un Peltier.

Chaque carte satellite contient les actionneurs et les circuits dédiés nécessaires à son contrôle. Le type de chaque satellite peut être identifié par le système hôte. Cette architecture matérielle offre une grande polyvalence et est facile à utiliser. À titre d'exemple, la figure 6.4 montre une configuration avec le système hôte enroulé autour du poignet connecté à trois satellites fournissant deux types de retours haptiques et un retour thermique.

Le retour haptique consiste en des vibrations dans une gamme de fréquences allant de 50 Hz à 550 Hz avec une résolution de 7 Hz. Quant à l'intensité de la vibration, elle doit être donnée en termes de pression sur la peau. Pour cela, il faut connaître la force mécanique appliquée par les actionneurs ; paramètre très difficile à obtenir car il dépend de la résistance variable de la peau de

l'utilisateur ainsi que de l'environnement (température, humidité relative, etc.). Les intensités de vibration sont donc données par la valeur relative utilisée dans nos protocoles pour la contrôler (0 = aucune vibration et 255 = vibration maximale). Le retour thermique consiste en des valeurs de température variant de +/- 5 ° C : la principale limitation est due au courant continu nécessaire qui peut réduire le temps d'expérience. Des variations jusqu'à +/- 10 ° C peuvent être facilement atteintes mais avec de sérieuses limitations de temps.

Avec le protocole ZigBee, jusqu'à 10 dispositifs hôtes peuvent être adressés. Et chaque dispositif hôte peut recevoir 8 cartes contrôlant jusqu'à 4 actionneurs. Finalement, jusqu'à 320 actionneurs pourraient être contrôlés au sein du dispositif TactiNET.

En résumé, le dispositif est conçu comme un prototype expérimental modulaire destiné aux chercheurs qui ne sont pas des experts en électronique. Ainsi, en fonction de l'objectif de l'étude, différentes combinaisons peuvent facilement être composées et évaluées, tant en terme de nombre d'actionneurs que de leur type. En outre, les actionneurs peuvent être librement échangés en mode *plug-and-play* et sont intégrés dans un boîtier en plastique fabriqué par une imprimante 3D. Chaque élément dispose d'un système d'accrochage afin de pouvoir être positionné sur différentes parties du corps de l'utilisateur. Dans nos expériences, décrites dans les sous-sections suivantes, un seul satellite est utilisé avec un retour haptique placé sur la main libre (non navigante).

6.2. Vers un langage graphique adapté à l'entrée textuelle et à la sortie vibrotactile

Dans l'univers du papier adapté aux non-voyants, les feuilles sont recouvertes de microcapsules qui gonflent à la chaleur, permettant de produire des dessins / schémas / graphiques en relief. La résolution particulière de la modalité tactile est prise en compte lorsqu'il s'agit de rendre accessible des compositions complexes. En atteste les recommandations telles que celles de la Figure 6.5 qui s'appuient sur une typologie de textures organisées en fonction de la possibilité de les discriminer les unes des autres.

Dans le cadre de l'application de notre métaphore de la canne blanche à l'amélioration de l'accessibilité des pages Web, il est rapidement advenu que la transposition analogique directe des contrastes lumineux de l'écran produisait une entrée sensorielle complexe et difficile à appréhender avec notre dispositif.

En s'inspirant des solutions proposées pour le papier thermogonflable, notre premier travail a consisté à conduire des protocoles expérimentaux et adapter notre métaphore comme indiqué par la figure 6.6. Accompagnés par des chercheurs en psychologie spécialistes de la modalité tactile, nous avons proposé :

— des règles de transposition de la structure visuelle des textes vers un langage graphique intermédiaire (basé sur des variations **contrôlées** de couleurs, de niveaux de gris et de formes) plus facile à appréhender avec notre dispositif (basé sur des variations de fréquence, d'intensité et de température) ;

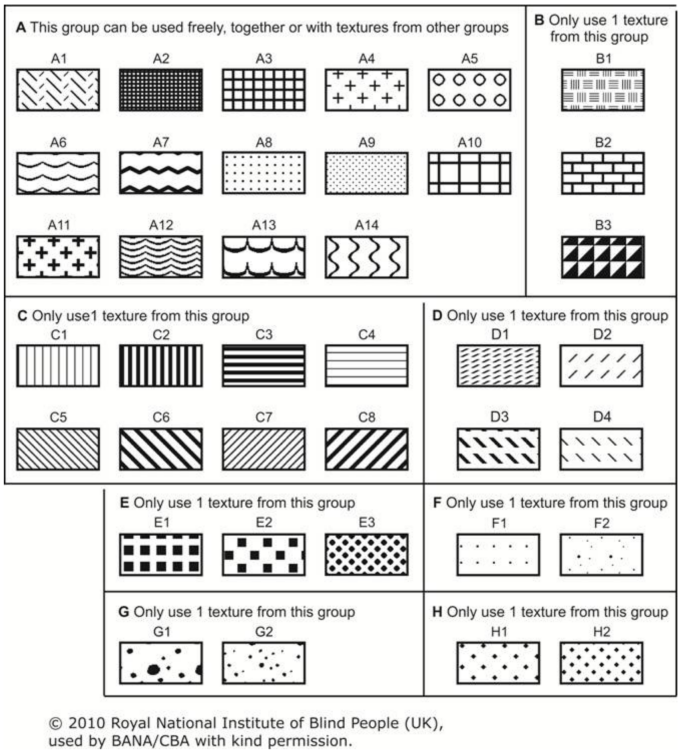


FIGURE 6.5. – Palette de textures pour papier à microcapsules.

— des règles de transposition analogiques de ce langage graphique en un langage composé de stimuli vibrotactiles et thermiques **discriminables** entre eux par un utilisateur non-voyant.

6.2.1. De la structure visuelle au langage graphique

Nous avons d'abord souhaité construire une preuve de concept en expérimentant le dispositif TactiNET pour reconnaître des formes simples représentant des mises en page web. Ce travail a également permis de fonder nos premiers choix de construction d'un langage graphique basé sur des motifs générant des stimuli variant en intensité et en fréquence.

La première expérience réalisée [106] visait à pré-tester la capacité des utilisateurs à reconnaître des formes dans un environnement non-visuel avec le dispositif TactiNET paramétré dans sa configuration minimale :

— un seul satellite positionné sur la main non active : un actionneur vibrotactile ;

— une seule dimension de variation : plus le pixel survolé par l'index est clair (respectivement foncé), plus l'amplitude de vibration est faible (respectivement élevée).

L'expérience a été menée avec 15 utilisateurs voyants (yeux fermés) et 5 personnes aveugles (voir tableau 6.1) pour explorer, compter, reconnaître et dessiner manuellement différentes configurations de pages web simulées. Les conclusions furent les suivantes. Premièrement, la capa-

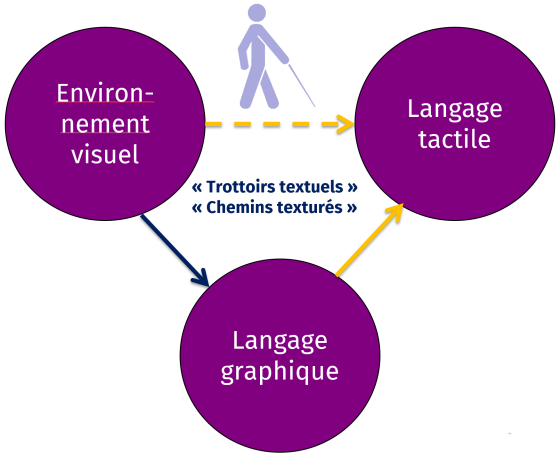


FIGURE 6.6. – Métaphore de la canne blanche adaptée au dispositif TactiNET.

Identifiant	1	2	3	4	5
Âge	63	67	59	56	36
Sexe	Homme	Femme	Homme	Femme	Femme
Âge de la cécité	0	32	25	10	15
Système d'exploitation	Linux	Linux	Windows	-	Windows
Technologie utilisée	ORCA	NVDA, ORCA	JAWS, NVDA	-	JAWS

TABLE 6.1. – Caractéristiques des usagers non-voyants soumis à l'expérimentation.

cité à distinguer la taille des formes et leurs relations spatiales a été évaluée comme très variable en termes de temps d'exploration (7 à 20 minutes au total pour explorer 4 configurations). La qualité des dessins manuels a également variée de très mauvaise à presque parfaite en fonction des caractéristiques de l'utilisateur (âge, précocité de la cécité, familiarité avec les technologies tactiles).

Cependant, comme le montre la figure 6.7, les meilleures productions sont qualitativement intéressantes malgré la configuration relativement simple du dispositif TactiNET. D'autres résultats secondaires intéressants méritent d'être notés. Premièrement, un effet d'apprentissage encourageant a été clairement mis en évidence alors que l'expérience n'a pas duré plus d'une heure. Deuxièmement, nous avons identifié une métrique permettant de mesurer l'intérêt de l'utilisateur pour l'information survolée : plus l'intérêt est grand, plus la pression exercée sur l'écran est proportionnellement appuyée.

Dans un second temps, afin de proposer un langage graphique capable de représenter plus efficacement les relations entre structure visuelle et motifs graphiques variables en formes, tailles, textures (de surface *vs.* de bordure), et espace d'encadrement, nous avons amélioré les capacités de transposition du dispositif en étudiant et en optimisant la combinaison de deux des différents stimuli vibrotactiles, à savoir l'amplitude et la fréquence.

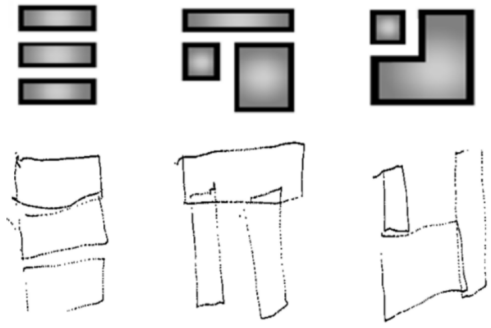


FIGURE 6.7. – Formes parcourues et exemple de dessins originaux reproduits par un utilisateur.

Dans cette perspective, nous avons conçu une deuxième expérience [138] destinée à sélectionner la gamme de fréquences la plus perceptible du dispositif TactiNET. Chaque valeur de fréquence étudiée a ensuite été combinée à une valeur d'amplitude soit constante, soit augmentée d'une légère variabilité aléatoire afin d'incorporer un effet de texture, susceptible de permettre des capacités perceptives plus importantes. Pour cela, nous avons mené une expérience qui a croisé 3 groupes d'utilisateurs (38 enfants voyants, 25 adultes voyants et 5 adultes aveugles) et qui consistait en une série de tests de comparaison sur un écran tactile divisé en deux zones distinctes. L'utilisateur devait décider si les stimuli perçus lors du survol de chaque zone étaient différents ou non. Tous les tests ont été réalisés avec un seul actionneur. L'utilisateur devait explorer la tablette avec le doigt d'une main, et avec l'actionneur placé sur l'autre main. La seule question posée après chaque exploration sans limite de temps était de savoir si les deux stimuli à gauche et à droite de l'interface étaient identiques. Initialement, toutes les valeurs d'amplitude étaient fixées, et seules les valeurs de fréquence variaient d'un stimulus à l'autre. Afin de minimiser les interférences et de maximiser la fluidité de l'expérience, une seconde tablette dédiée à l'expérimentateur était reliée à la première par une connexion *Bluetooth*. Elle était équipée d'une interface permettant de contrôler rapidement et à distance la présentation et la valeur successive des séries de stimuli. Chaque série était composée d'une valeur de référence fixe envoyée sur le côté gauche de la tablette et d'une valeur de comparaison variable envoyée sur le côté droit. Le principal résultat indique que les participants étaient collectivement plus sensibles aux différences de fréquences proches de 300 Hz. Cette capacité perceptive s'individualise et se détériore à mesure que l'on s'éloigne de cette valeur. Ce résultat peut être comparé à [34, 156], qui indiquent que la navigation tactile sur les appareils numériques est plus sensible aux vibrations dont les fréquences sont autour de 250 Hz. D'autre part, l'expérience n'a pas montré d'effet significatif de l'amplitude-bruit, de l'âge ou de la cécité (sauf pour les enfants qui sont significativement plus sensibles aux séries ascendantes que descendantes). En résumé, les conclusions concernant les premières règles grammaticales applicables à notre langage graphique se déclinent selon les 5 plages de fréquences énumérées ci-dessous :

— entre 50 Hz et 150 Hz, le seuil perceptif minimum est de 15 Hz ;

— entre 150 Hz et 250 Hz, le seuil perceptif minimum est de 13 Hz ;

— entre 250 Hz et 350 Hz, le seuil perceptif minimum est de 7 Hz ;

— entre 350 Hz et 450 Hz, le seuil perceptif minimum est de 10 Hz ;

— entre 450 Hz et 550 Hz, le seuil perceptif minimum est de 15 Hz.

Ces valeurs de fréquence furent ensuite combinées avec des valeurs d'amplitude afin d'augmenter le pouvoir expressif du langage graphique. Par conséquent, nous avons conçu une troisième expérience [139] pour sélectionner la gamme d'amplitude la plus perceptible. Chaque valeur d'amplitude étudiée a ensuite été combinée avec la valeur de fréquence optimisée obtenue lors de l'expérience précédente. Le protocole était très similaire au précédent, à l'exception de la population (20 adultes voyants et 5 adultes aveugles), des valeurs de référence de l'amplitude comprises entre 55 et 255, et de la méthode utilisée pour déterminer le seuil de perception de l'amplitude [29]. De nombreuses valeurs ont été comparées à trois amplitudes de référence en croisant les deux populations d'utilisateurs. Les meilleurs résultats se situent dans un intervalle autour de la valeur 55 quelle que soit la condition expérimentale. Néanmoins, bien que mesurant des seuils perceptifs moins sensibles, les valeurs les plus élevées révèlent une différence significative entre les groupes de voyants et d'aveugles.

Sur la base de ces résultats, nous avons augmenté notre grammaire des 3 règles suivantes :

— entre 0 et 100, le seuil perceptif minimum est de 12 ;

— entre 100 et 200, le seuil perceptif minimum est de 48 ;

— entre 200 et 255, le seuil perceptif minimum est de 45 ;

A travers ces expériences, nous avons commencé à tisser des liens entre un langage graphique et les paramètres de base des stimuli vibrotactiles. Dans notre dernière expérience, présentée dans la sous-section suivante, nous tentons de mettre en relation ces résultats avec les structures visuelles extraites d'un corpus de 900 pages Web.

6.2.2. Reconnaissance des structures de pages Web

Afin de tester les possibilités du dispositif TactiNET, nous avons conduit une expérience exploratoire basée sur la transposition semi-automatique des structures visuelles des pages Web en stimuli vibrotactiles. Cette conversion est réalisée en divisant une page Web donnée en zones d'intérêt représentées par des objets graphiques rectangulaires et dont les contrastes formels sont ensuite directement traduits en stimuli vibrotactiles.

Hypothèses et *design*

Après avoir sélectionné des pages Web représentatives de notre contexte applicatif (sites de e-commerce, de tourisme, d'information), un algorithme de partitionnement basé sur des graphes agglomératifs décrit dans [137] a été appliqué à chacune d'entre elles pour obtenir leur découpage en un nombre demandé de zones (figure 6.8). Pour cette expérience, nous avons fixé le nombre de zones à 5. Ce nombre provient de [110], qui a montré l'existence d'une limite en termes de quantité d'informations qu'une personne peut recevoir, traiter et retenir. La théorie, connue sous le nom de loi de Miller, propose que le nombre moyen d'objets qu'une personne peut conserver en mémoire de travail est de 7 +/- 2. Ainsi, pour garantir que le nombre de

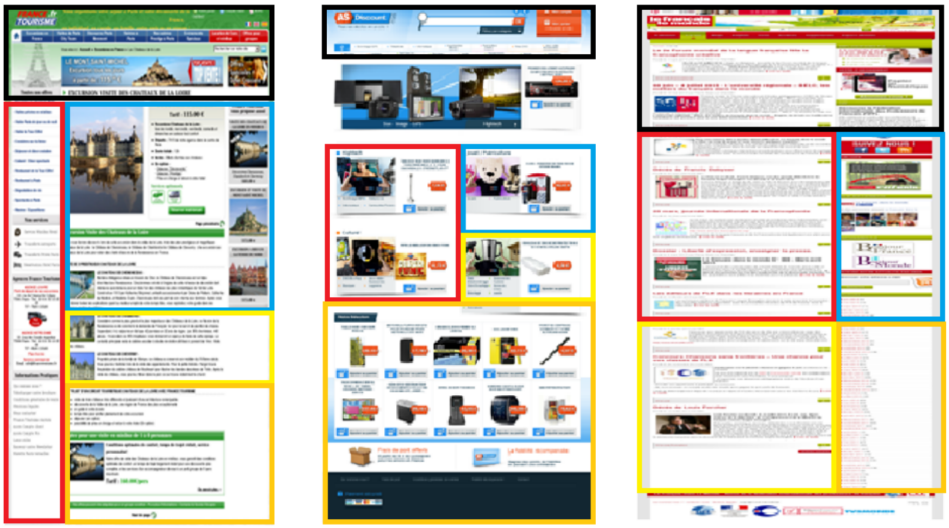


FIGURE 6.8. – Partitionnement de pages Web représentatives de nos trois domaines d'application (tourisme, e-commerce, information).

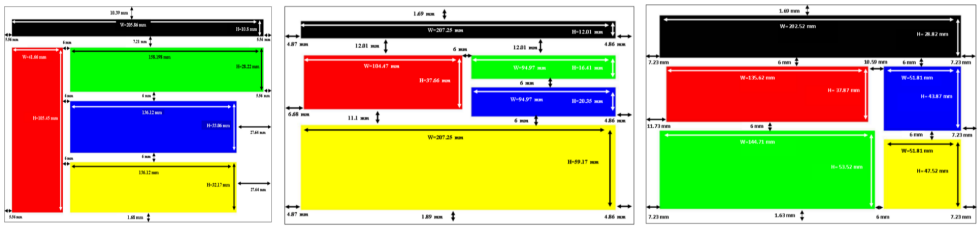


FIGURE 6.9. – Segmentation adaptée aux conditions expérimentales.

partitions choisi n'affecte pas la performance des participants, nous avons fixé cette valeur au minimum de la capacité. Les résultats du partitionnement ont ensuite été adaptés à notre configuration expérimentale pour prendre en compte les recommandations pour la navigation non visuelle sur les dispositifs à écran tactile (figure 6.9). Ensuite, sur la base des résultats présentés dans la section précédente en termes de seuils de perception, une valeur d'écart-type représentant le contraste visuel était associée à chaque zone proportionnellement à une valeur d'amplitude, tandis qu'une valeur constante et optimale était attribuée à la fréquence : plus le contraste visuel était important, plus la variance colorimétrique des pixels était élevée, plus la force du stimulus tactile était importante. Une page vibrante de base a également été construite pour servir de référence en fixant l'amplitude pour toutes les zones à la valeur optimale établie dans nos précédentes expériences.

En optimisant à la fois les paramètres vibrotactiles et la différence dans les structures visuelles générées, l'expérience consistait à évaluer deux hypothèses : (H1) le TactiNET augmente la capacité à discerner les différentes catégories de pages vibrantes (c'est-à-dire si deux pages découvertes sont similaires ou différentes) et (H2) de meilleures performances dans cette tâche

sont obtenues en utilisant des valeurs d'amplitudes variables.

FIGURE 6.10. – Comparaison des structures de pages Web en fonction des valeurs d'amplitude associées.

Au total, 36 comparaisons de structures de pages Web ont été effectuées par chacun des 11 participants (6 non-voyants et 5 voyants). Il y a 3 comparaisons de structures identiques et 3 comparaisons de structures de pages différentes. Toutes les comparaisons sont répétées 3 fois pour éviter une sélection aléatoire. Et, toutes les comparaisons sont soumises à 2 conditions d'amplitude : variable ou fixe. Dans les mêmes conditions que l'expérience précédente, chaque tâche comprend une série de tests, qui présentent les structures à comparer sur deux tablettes à écran tactile (figure 6.10). Après la phase d'exploration, le participant décide si les deux pages vibrantes sont identiques ou non. Enfin, à la fin des 36 comparaisons, l'utilisateur devait dessiner la dernière structure découverte sur une feuille de papier A4.

Principaux résultats

Plusieurs remarques générales peuvent être observées. Tout d'abord, les résultats sont globalement homogènes entre les participants voyants (yeux fermés) et non-voyants, quelle que soit la condition d'amplitude, avec un ratio d'environ 60% de bonnes réponses. Ce score de performance semble prometteur car (1) 5 aveugles sur 6 n'utilisent jamais de dispositifs tactiles, (2) aucun d'entre eux n'a eu de temps d'entraînement et l'expérience a été longue (en moyenne 1h30) et ennuyeuse, et (3) le dispositif était dans sa configuration minimale c'est-à-dire un seul actionneur et une seule dimension de variation (l'amplitude). Nous notons tout de même une supériorité pour la population aveugle dans le cas de l'amplitude variable. Cette remarque est importante en particulier car cette supériorité provient exclusivement du cas des comparaisons entre structures similaires. En effet, Si les résultats confirment logiquement qu'il est deux fois plus facile d'identifier une dissemblance entre deux structures vibrotactiles que d'être certain de leur identicité, **les aveugles semblent plus aptes à exploiter la variation d'amplitude pour lever leurs doutes sur la similitude de deux structures**.

Quelques remarques concernent la qualité des dessins produits à la fin de l'expérience, car aucune information préalable n'avait été donnée sur la nature des éléments explorés. Les voyants (yeux fermés) étaient plus à l'aise pour représenter naturellement leur perception sous forme de rectangles, mais le dessin était souvent plus complexe que la structure explorée. En parallèle,

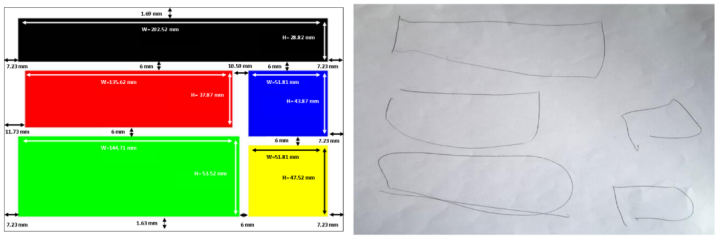


FIGURE 6.11. – Exemple de reproduction par un non-voyant de la page Web information.

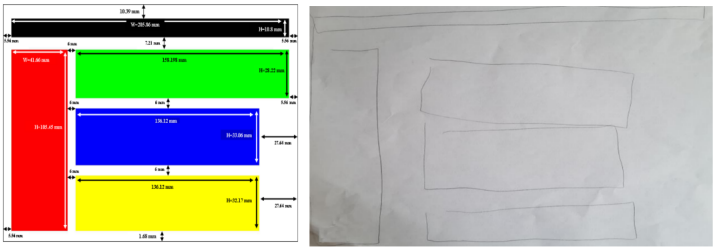


FIGURE 6.12. – Exemple de reproduction par un non-voyant de la page web tourisme.

trois des personnes non voyantes n'ont pas pu abstraire cette notion et ont reproduit le chemin suivi par leur doigt sous forme de lignes. En revanche, **deux aveugles ont reproduit très fidèlement les structures, alors même qu'aucune indication formelle n'avait été donnée** sur le support du stimulus vibratoire (voir figures 6.11 et 6.12). Dans tous les cas, cet exercice nous a semblé être une tâche intéressante, même avec une population d'utilisateurs aveugles, pour évaluer la capacité de nos stimuli à construire une représentation mentale de la structure visuelle. Nous observons d'ailleurs **pour les non-voyants une tendance de corrélation linéaire entre les scores donnés aux dessins par des experts et les score de performance de la tâche de comparaison**.

Toutes les expériences ayant été filmées, nous avons pu observer les différentes stratégies utilisées par les participants pour explorer l'écran tactile de la tablette. Une première typologie a pu être dégagée. Ce qui nous semble remarquable est la richesse des micro-stratégies utilisées et donc le grand nombre de combinaisons possibles pour les ancrer dans une stratégie globale. Cette expérience ne nous permet pas encore de définir une relation claire entre les combinaisons de micro-stratégies et l'efficacité globale pour la reconnaissance de structures textuelles visuelles. Néanmoins, l'analogie avec les capacités visuelles préperceptives semble suffisamment évidente pour étendre cette axe de recherche : poser une base solide pour l'exploitation des processus de *scanning* non visuels et donc le développement de stratégies de lecture tactile efficaces. Une analyse préliminaire peut être produite à partir des trois observations suivantes :

— la majorité des participants commencent à naviguer sur les tablettes de gauche à droite. Cela pourrait être dû aux habitudes culturelles liées aux directions d'écriture et de lecture ;

— les participants utilisent entre 2 et 6 micro-stratégies, qui peuvent être classées en trois catégories : (1) navigation continue prenant des informations dans les deux dimensions

horizontale et verticale, (2) navigation utilisant une seule direction (horizontale, verticale ou diagonale) et (3) navigation n'utilisant aucune de ces possibilités.

— deux micro-stratégies particulières de type boustrophédon (alternance du sens de lecture d'une ligne à l'autre) sont utilisées par tous les participants (une stratégie horizontale et une stratégie verticale).

— il est à noter des différences remarquables entre les participants dans le temps et la vitesse d'exécution de la demande en fonction des micro-stratégies utilisées.

Ces remarques nous amènent à formuler plusieurs hypothèses. La structure survolée n'étant pas connue à l'avance, les stratégies naturelles mises en œuvre sont préférentiellement basées sur un parcours horizontal et vertical continu de l'ensemble de l'écran. Or, l'efficacité de la capture d'information est dégradée par l'enchaînement de micro-stratégies trop nombreuses ou trop différentes. Ces résultats participeraient à expliquer les trois scores les plus faibles des personnes aveugles qui on mis en place de nombreuses stratégies différentes. **Les stratégies les plus efficaces ont en commun une reconnaissance en seulement deux étapes : l'information est d'abord prise par un parcours continu dans les deux dimensions horizontale et verticale de l'écran, puis suivie d'une vérification, essentiellement verticale, pour lever certaines incertitudes**.

En examinant les résultats détaillés par participant ou par catégorie de structures visuelles, quelques éléments de discussion apparaissent également. Tout d'abord, seule la structure de page Web « tourisme » explorée par les aveugles semble influencer les scores ; c'est-à-dire qu'ils sont plus élevés pour les aveugles que pour les participants voyants. Toutes les comparaisons impliquant cette structure visuelle entraînent systématiquement (à une exception près) un score meilleur pour les non-voyants, aussi bien pour découvrir une similarité qu'une différente. L'explication que nous proposons se trouve dans la congruence entre les formes survolées et la valeur d'amplitude choisie. En effet, la catégorie tourisme est la seule pour laquelle les valeurs d'intensité sont proportionnelles aux valeurs de taille/surface des zones qui y sont associées. Cette conclusion, si elle est prouvée par d'autres expériences sur une vaste population, est très intéressante pour soutenir notre approche car **elle légitime l'utilisation de critères objectifs (ici variance de contraste), extraits du document source et transformés de manière analogique, pour construire un paysage tactile cohérent**.

Enfin, il semble y avoir une tendance à ce que les scores soient d'autant plus élevés que l'apparition de la cécité est précoce. Outre les conditions d'apparition de la cécité, nous noterons dans nos résultats une tendance à une tâche plus facile pour les aveugles qui sont familiers avec les nouvelles technologies tactiles, et pour ceux qui passent beaucoup de temps à surfer sur le Web. D'ailleurs, **la meilleure performance est attribuable à la seule femme aveugle du protocole (avec près de 90% de bonnes réponses dans toutes les conditions), qui est la plus jeune, possède un *iPhone* équipé de *VoiceOver*, et est connectée plus de 10 heures par jour**.

6.2.3. Bilan

Dans ce travail, nous avons développé un dispositif vibrotactile appelé TactiNET pour l'exploration active de la mise en page et de la typographie de pages web dans un environnement

non-visuel ; l'idée étant d'accéder à la sémantique morpho-dispositionnelle du message véhiculée par l'architecture textuelle. Pour cela, nous avons d'abord construit un dispositif expérimental permettant la transposition analogique des contrastes lumineux émis par une tablette tactile en stimuli vibratoires et thermiques. Les principales contraintes ergonomiques étaient d'être facilement positionnable sur n'importe quelle partie du corps, modulable en terme de qualité et de quantité d'actionneurs, peu coûteux, robuste et léger. Nous avons ensuite réglé le dispositif pour poser les briques d'un langage tactile en fonction de la capacité expressive des stimuli produits. Ce travail a été initié par l'étude des seuils minimaux de perception de la fréquence et de l'amplitude de la vibration. Enfin, nous avons évalué la capacité du dispositif TactiNET à permettre la catégorisation de pages Web de trois domaines, à savoir le tourisme, le commerce électronique et les actualités, présentées par une adaptation vibrotactile de leur structure visuelle.

Bien qu'exploratoires, les expériences sont particulièrement encourageantes, renforcées par le fait que nous avons délibérément choisi des conditions très défavorables, c'est-à-dire (1) l'hétérogénéité de la population aveugle en termes d'âge, d'habituation aux technologies tactiles et à la navigation Web, de début de cécité, et (2) la configuration minimale de notre dispositif avec un seul actionneur vibrotactile et une seule dimension de variation. Malgré cela, les hypothèses intéressantes que nous retenons sont les suivantes :

— les aveugles ont tendance à affirmer la similitude entre deux structures mieux que les voyants, surtout lorsque les relations entre la forme survolée et les intensités perçues sont cohérentes ;

— les aveugles semblent être capables d'imaginer les formes qu'ils ont ressenties sans aucune indication préalable des stimuli ;

— les utilisateurs semblent développer un riche ensemble de micro-stratégies pour naviguer sur l'écran tactile vibrant ;

— l'utilisation régulière des technologies tactiles et le nombre d'heures quotidiennes passées sur le web semblent être un facteur positif pour l'appropriation du dispositif.

Cette première étude exploratoire ouvre un grand nombre de pistes de recherche. Par exemple, une direction possible est d'affiner sur des critères de pression, spatiaux et temporels la typologie des micro-stratégies d'exploration des formes. L'objectif sera d'analyser les relations entre les micro-stratégies et les macro-stratégies d'accès à l'information. Nous pensons qu'à travers cette alternance interactive entre perceptions locales et globales, nos capacités naturelles de *skimming* et de *skimming* peuvent être exploitées, que ce soit dans la modalité visuelle ou tactile. Une des perspectives principales de ce travail est également d'explorer expérimentalement des configurations plus complexes de notre dispositif et ainsi d'améliorer l'expressivité de notre langage tactile. Dans un premier temps, nous proposerons d'étudier l'association des paramètres visuels et thermiques. Deuxièmement, des travaux sont en cours pour évaluer la possibilité de combiner des stimuli tactiles avec des stimuli audio [103]. Troisièmement, l'augmentation du nombre de stimuli simultanés (jusqu'à 320) en associant un actionneur vibrotactile à plusieurs doigts pourrait permettre de nouvelles expériences sensorielles.

Enfin, nous noterons l'émergence dans notre expérience et par des discussions informelles avec des personnes aveugles d'une fonctionnalité inattendue. En effet, une des difficultés fréquentes rencontrées par les personnes aveugles réside dans la possibilité d'avoir une idée *a priori*

de la taille d'un document. Dans nos expériences, certaines des formes choisies pour calibrer le dispositif présentaient une gradation du noir au blanc sur toute leur surface (c'est-à-dire que plus le niveau de gris était élevé, plus la vibration de l'actionneur était forte). Par conséquent, la gradation produisait une vibration continuellement décroissante (ou croissante) lorsque le doigt la survolait. De nombreux aveugles ont rapporté avoir ressenti des différences dans la vitesse de cette diminution dès le début de leur exploration de la forme : la finesse du dégradée était d'autant plus affirmée, et donc la transition du noir vers le blanc d'autant plus lente, que la forme survolée était étendue. En fait, nous avions construit par hasard un stimulus qui permettait aux aveugles d'anticiper la taille d'une zone. Il ne semblait pas nécessaire à certains participants de la survoler plus de quelques centimètres pour interpréter cette propriété et anticiper le temps de parcours complet de sa forme.

7. Application au projet TagThunder : métaphore de la *cocktail party*

Le travail décrit dans les sections de ce chapitre a été initié par deux Contrats Plan Etat Région (CPER) au sein du programme interdisciplinaire transversal SHS/STIC NUMNIE, puis développé sous ma responsabilité scientifique dans le cadre de l'appel à projet du Fonds National pour la Société Numérique dédié à l'Accessibilité Numérique. Le projet lauréat **TagThunder** a pu être mis en place en partenariat avec l'entreprise *AccessMan* et financé par la Banque Publique d'Investissement dans le cadre du Plan d'Investissement Avenir. Une publication scientifique dans une revue internationale [73] détaille les principaux tenants et aboutissants de cette recherche ; une valorisation du projet et de ses résultats peut être visionnée sur *https://www.youtube.com/watch?v=vrET36OcjJs&t=124s* et testée sur *https://tagthunder.greyc.fr/demo/*.

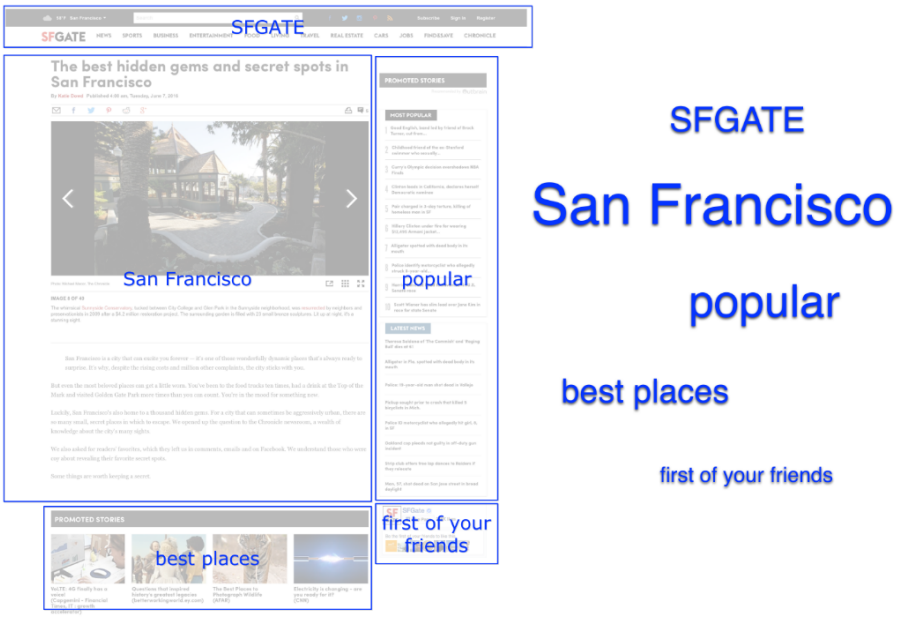


FIGURE 7.1. – De la page Web au nuage de mots.

Observons la page web à gauche de la figure 7.1. Poursuivons avec une segmentation en zones et l'extraction, pour chaque zone, de quelques mots-clés représentatifs du contenu (figure 7.1, gauche). Enfin, effaçons les autres éléments de la page ; disposons les termes retenus dans la même relation spatiale que la zone qu'ils représentent et façonnons-les graphiquement pour les rendre d'autant plus saillants que la zone semble importante. Nous obtenons un stimulus visuel fréquemment utilisé dans le web 2.0 : un nuage de mots (figure 7.1, droite). L'idée est de transposer ce concept dans le monde du son, en transformant ce nuage de mots en un « tonnerre » de mots (ou tag thunder). L'analogie qui sous-tend le développement de ce concept est une extension de la métaphore connue des psychologues sous le nom de « *cocktail party effect* » [88].

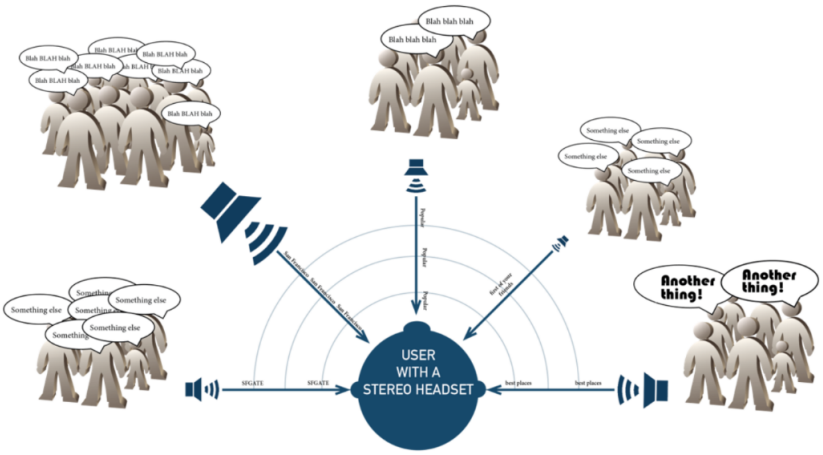


FIGURE 7.2. – Métaphore de la *cocktail party*.

En psycho-acoustique, cette métaphore désigne la possibilité de focaliser son attention auditive sur un flux verbal dans l'atmosphère bruyante d'une réception, qu'elle soit dirigée vers des sources sonores extérieures à la conversation ou vers ses interlocuteurs. Nous suivrons ce fil métaphorique en considérant la relation entre le lecteur aveugle et les zones de la page visitée, de la même manière qu'entre un invité, situé au centre d'une salle, et les différents groupes de discussion qui s'y sont formés : les échanges sont séquentiels au sein d'un groupe mais concurrents entre les différents groupes ; l'invité, en tant que nouveau venu, doit prendre suffisamment d'informations dans l'environnement sonore pour identifier la discussion dans laquelle il souhaite s'impliquer. Par exemple, dans la figure 7.2, certains termes émis par chaque groupe de personnes sont captés par l'auditeur comme des flux répétés provenant de sources parallèles distinctes. La force, la densité et la vitesse des flux dépendent des caractéristiques du groupe émetteur ; la métaphore suggère donc une relation entre les caractéristiques formelles de la zone et des paramètres acoustiques spécifiques.

Une des solutions qui a été évoquée dans plusieurs recherches pour améliorer la vitesse de lecture des personnes aveugles, est basée sur la notion de parole simultanée. En particulier [65] montre expérimentalement l'intérêt de cette approche mais aussi certaines limites percep-

tives. Pour aller plus loin, une des originalités de notre approche est de proposer d'exploiter la spatialisation sonore 3D et les techniques d'audition binaurale pour surmonter ces difficultés en optimisant la séparation perceptive des sources sonores [69]. Dans le travail présenté ici, la spatialisation est limitée à un demi-plan, perpendiculaire à l'utilisateur, dans lequel les 5 sources sonores sont placées devant l'auditeur sur la gauche, la droite, le centre et les deux diagonales. Les sections suivantes décrivent la plate-forme expérimentale et l'architecture logicielle qui sous-tend son fonctionnement, les résultats des premières preuves de concept des interactions homme-machine qu'elle permet. Le travail effectué plus en profondeur sur certains algorithmes issues des domaines de l'apprentissage machine et du traitement automatique des langues pour réaliser la phase de segmentation automatique de pages Web (nécessaire aussi bien au projet TactiNET que Tagthunder) sera présenté dans le chapitre 8.

7.1. Architecture logicielle

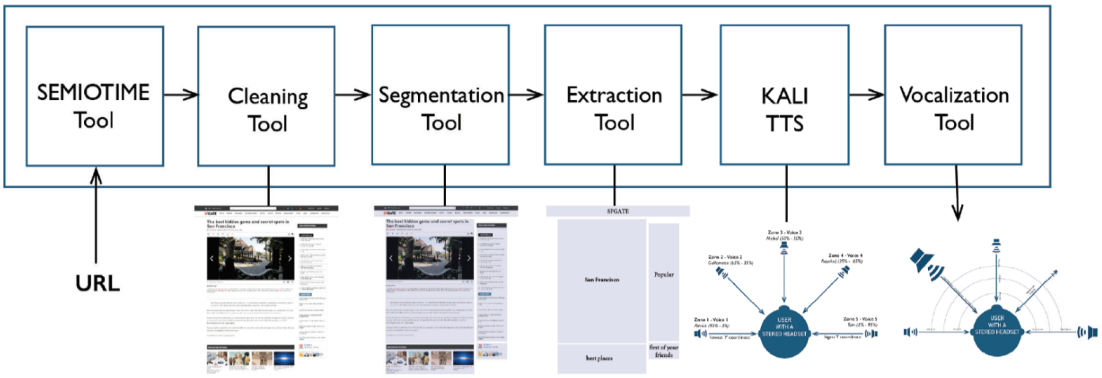


FIGURE 7.3. – Architecture modulaire : de l'*URL* au tonnerre de mots.

Pour créer un tonnerre de mots à partir d'une *URL*, outre les opérations techniques de centralisation des informations à impact visuel (injection dynamique de *JavaScript* pour regrouper et sélectionner toutes les propriétés de style utiles), notre solution repose sur un enchaînement de différentes couches logicielles. En particulier, les opérations de segmentation des pages Web, d'extraction d'expressions-clés et de génération orale des nuages de mots sont gérées par un *Web service* organisé en *pipeline* et interrogé par une extension du navigateur *Firefox*. La figure 7.3 illustre cette architecture et ses différents modules, dont les 3 principaux sont rapidement décrits dans les sous-sections suivantes en terme d'approche théorique, d'état de développement et d'évaluation.

7.1.1. Segmentation de pages Web

Notre approche dirigée par la tâche pose un certain nombre de contraintes d'ordre général. Selon notre analyse, le *skimming* non visuel impose trois pré-requis à nos propositions algorith-

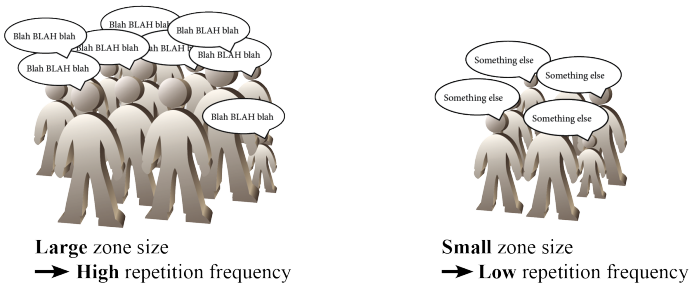
miques.

Tout d'abord, le nombre de zones doit être fixé en amont afin de (1) favoriser l'émergence de régularités dans la sortie sonore et (2) respecter le nombre maximal de stimuli oraux simultanés qu'un être humain peut distinguer cognitivement. Dans ce contexte, nous avons montré que la charge cognitive peut s'élever jusqu'à 5 stimuli différents, limitant ainsi le nombre maximum de zones résultant du processus de segmentation de pages Web. A moyen terme, la stabilité de la segmentation de la page Web en 5 zones exactement aidera l'utilisateur à construire des associations entre la position spatiale d'un son et une fonction logique souvent observée au même endroit pour une catégorie donnée de pages Web (en-tête, panier d'achat, contact, ...). Deuxièmement, chaque zone doit être associée à une source sonore unique située dans l'espace en fonction de sa position dans la page web. Ainsi, chaque zone doit être un bloc compact composé d'éléments web contigus ; pour la même raison, les zones ne doivent pas non plus se chevaucher. Troisièmement, la segmentation doit être complète, ce qui signifie qu'aucun élément visible de la page Web ne doit rester en dehors d'une zone donnée, car l'objectif est de révéler la structure visuelle globale d'un document sans *a priori* d'importance ou tri de l'information contenue.

Ce module est celui le plus étudié dans nos différents travaux. Plusieurs stagiaires, ingénieurs, doctorants et post-doctorants ont permis de contribuer à l'objectif de segmentation automatique de pages Web. Cela nous a suggéré de renvoyer le lecteur à une section plus complète dédiée à ce thème (section 8).

7.1.2. Extraction d'expressions-clés

L'extraction des expressions-clés est réalisée à l'aide d'un algorithme classique basé sur TF-IDF (Term Frequency-Inverse Document Frequency). La mesure TF-IDF, utilisée dans les moteurs de recherche, trouve des n-grammes spécifiques au contexte puisque les fréquences des termes sont pondérées par l'inverse de leur fréquence d'apparition dans un corpus. Inspirée de [166], pour appuyer notre approche basée sur l'intégration des propriétés de structure visuelle dans nos algorithmes, notre solution complète l'extraction par une pondération des scores de TF-IDF par la position des termes dans les blocs de texte ou encore par leur mise en forme. Ce principe a été amélioré par [74], en utilisant un *ranking SVM* au lieu de la position brute. Le corpus utilisé pour calculer la fréquence inverse des documents (IDF) est constitué de 953 551 articles du journal « Le Monde », couvrant 20 ans, de 1987 à 2006. Les données les plus récentes du corpus datant de 2006, elles ont introduit un certain silence et un biais dans le processus d'extraction des expressions-clés. Bien qu'elle n'ait pas encore été évaluée, cette technique semble fournir des premiers résultats limités mais satisfaisants uniquement lorsque les zones contiennent suffisamment de texte. Cette dernière difficulté, abordée dans les perspectives, sera questionnée en travaillant sur l'intégration d'informations méta-textuelles issues de la description des zones ou des images qui les composent.



Filage métaphorique 1 : plus le groupe qui parle d'un sujet est imposant, plus les termes connexes à la thématique de la discussion émergent souvent.
Analogie métaphorique 1 : les expressions-clés vocalisées sont jouées en boucle. La taille de la zone influence leur fréquence de répétition à l'intérieur de la boucle.

FIGURE 7.4. – Métaphore pour la fréquence de répétition

7.1.3. Spatialisation sonore

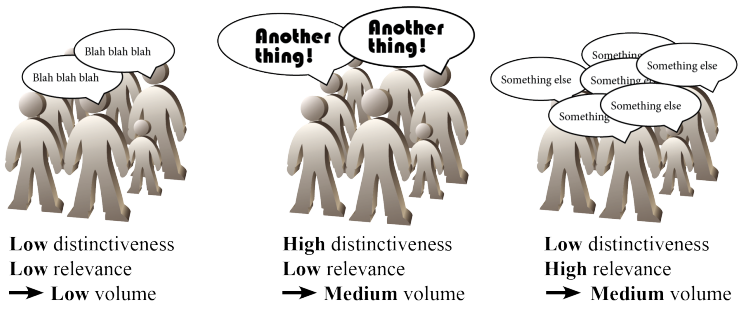
En s'appuyant dur des études telles que [21, 42, 158], la voix spécifique, le volume, la prosodie, le débit de parole et la synchronisation du son sont combinés pour générer le signal audio à partir d'une expression-clé donnée par l'étape précédente d'extraction et les propriétés visuelles de la zone qui lui est associée.

Notre module de synthèse utilise l'outil Kali TTS [111], développé par le laboratoire CRISCO à Caen. Kali supporte l'accélération du débit de parole sans perte d'intelligibilité et de qualité sonore, ce qui est une caractéristique très importante dans la navigation Web non visuelle. Nous filons de différentes manières la métaphore de la *cocktail party* pour attribuer aux expressions-clés synthétisées une fréquence de répétition, un volume et un emplacement dans l'espace audio. La vocalisation de toutes les expressions-clés avec leurs paramètres spécifiques produit ainsi le tonnerre de mots final retourné par le *Web service* au client Web. Dans nos premières expérimentations nous avons proposé un filage métaphorique qui génère les règles exposées ci-après et illustrées par les figures 7.4, 7.5 et 7.6.

— Pour chaque expression-clé, le silence entre deux répétitions dans la boucle est inversement proportionnel à la taille relative de sa zone d'appartenance. Plus la zone est grande, plus le silence est court. (Figure 7.4).

— Pour chaque zone, le valeur en contraste d'une zone est calculé en fonction du rapport entre la variance en contraste de la page Web et celle de la zone. Le volume est défini en utilisant cette valeur et la fréquence de répétition de l'expression-clé dans la zone (Figure 7.5).

— Les voix sont réparties de manière égale sur 5 positions de l'espace stéréo 2D en fonction des coordonnées du centroïde de la zone, comme l'illustre la figure 7.6.



Filage métaphorique 2.a (caractère distinctif) : plus une voix se distingue dans un groupe, plus il est facile de détecter sa source.
Filage métaphorique 2.b (pertinence) : plus les mots sont répétés dans un groupe, plus ils sont pertinents.
Analogie métaphorique 2 : la calcul de la valeur de volume d'une source sonore implique la force du contraste de la zone et la fréquence de répétition de l'expression-clé dans son contenu textuel.

FIGURE 7.5. – Métaphore pour le volume

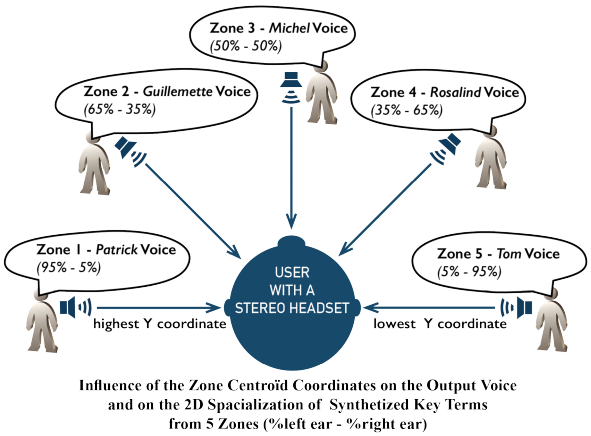
7.2. Preuve de concept

Dans l'objectif d'acquérir nos premièrs retours sur la plausibilité du concept de Tagthunder, une première expérience a été réalisée avec des utilisateurs voyants selon le protocole suivant détaillé dans [83]. L'objectif principal était d'explorer à quel point un tel stimulus avait la capacité de représenter rapidement la structure logico-thématique de pages Web variables en longueur, densité textuelle ou structuration visuelle.

Dans ce cadre, les termes-clés susceptibles de composer les Tagthunders ont été extraits des zones de texte sous la forme de n-grammes de mots (sous-séquences de n mots construites à partir d'une séquence textuelle donnée) ; nous avons limité la longueur des séquences à n < 7 mots. Afin d'associer un score à chaque n-gramme, nous avons classiquement calculé un tf-idf [151] (*term frequency - inverse document frequency*). Le tf correspond à la fréquence d'un terme donné dans une zone ; et l'idf est calculée à partir d'un corpus contenant 953 551 articles du journal « Le Monde » de 1987 à 2006. Comme pour [166], notre solution a couplé la métrique tf-idf avec quelques paramètres supplémentaires composant la formule (7.1) utilisée pour calculer le score final de chaque terme-clé. Pour les besoins de notre expérience, chaque zone n'est représentée que par un seul terme-clé.

$$Score = tf(terme, zone) \cdot idf(terme, C) \cdot \sum_{i=1}^{n} \sigma(c_i) \tag{7.1}$$

où $tf(terme, zone)$ est la fréquence du terme dans sa zone, $idf(terme, C)$ est le nombre de documents de notre corpus C contenant le terme. $\sigma(c_i)$ est le poids d'une caractéristique c_i telle que le poids, la taille, la variante, le style de la police, etc. Les valeurs σ ont été attribuées de



Filage métaphorique (dimensions spatiales) 3 : la spatialisation du son permet de positionner et distinguer plusieurs groupes de discussion.
Analogie métaphorique 3 : Les coordonnées des zones influencent le type de voix de sortie et la spatialisation en 2D des mots-clés vocalisés.

FIGURE 7.6. – Métaphore pour le positionnement des sources dans l'espace

manière empirique et reflètent la perception visuelle. Par exemple, un terme en début de zone est considéré comme plus important, les termes clés à l'intérieur d'un titre sont davantage pondérés qu'à l'intérieur d'un paragraphe ; une hauteur de ligne plus large appuie également la valeur des termes qui la compose.

L'objectif de cette expérience est double : (1) évaluer la pertinence des termes clés extraits et (2) tester l'efficacité du concept de Tagthunder comme stratégie de prise rapide d'information (*skimming*).

L'expérience proprement dite s'est déroulé comme suit. Un participant voit un **nuage** de mots suivi d'une page Web, 15 secondes chacun. La page peut être ou non la page Web dont les mots ont été réellement extraits. Il est demandé au participant si le nuage de mots correspond selon lui à la page affichée parmi les réponses possibles : oui, plutôt oui, plutôt non, non. Les mêmes données sont présentées à un autre participant, mais sous la forme d'un **tonnerre** de mots, donc à écouter au lieu du nuage de mots à voir ; le participant est invité à répondre à la même question d'adéquation entre le stimulus audio et la page Web explorée pendant 15 secondes après l'écoute.

Les modalités de l'expérience étaient les suivantes :

— 18 participants voyants, chacun avec 16 stimuli différents (8 nuages de mots - 8 tonnerres de mots) ;

— 24 pages Web provenant de divers sites Web ont été utilisées pour générer un nuage de mots et un tonnerre de mots pour chaque page ;

— 24 autres pages Web ont été sélectionnées pour créer des stimuli qui ne correspondent pas ;

FIGURE 7.7. – Interface Web d'expérimentation

Chaque paire (page Web, nuage/tonnerre de mots) a été montré à 3 participants différents ; il y avait autant de paires correctes (en adéquation) que de paires incorrectes. Les participants ont effectué le test de manière autonome, avec un superviseur à proximité, comme le montre l'illustration de la plate-forme d'expérimentation conçue à cet effet (Figure 7.7).

De manière générale, les résultats, détaillés dans [83], montrent que les participants ont en même temps trouvé l'exercice difficile mais fait très peu d'erreurs d'appariement. De plus, les performances relevées pour la tâche impliquant les **tonnerres** de mots sont comparables, en termes de précision globale, à celle impliquant des *nuages* de mots. Nous avons conclu des performances objectives que le concept de tonnerre de mots est intéressant mais que les avis subjectifs des participants soulignent certaines limitations provenant en partie de la qualité des modules précédents la vocalisation. Il reste ici toute la difficulté de construire une expérimentation pour simultanément (1) évaluer un système modulaire dans son ensemble au bout de la chaîne de traitement et (2) identifier au niveau de chaque brique logicielle l'origine exacte de ses limites.

7.3. Bilan

Dans cette partie, nous avons présenté une approche générale et théorique du problème de l'accès rapide, global et non visuel à la navigation Web. Nous sommes partis du constat que la sémantique de l'architecture visuelle des pages Web doit être transposée dans de nouvelles modalités sensorielles. Pour permettre aux utilisateurs aveugles d'exploiter ces nouveaux stimuli et augmenter leur capacité à développer des stratégies de lecture de haut niveau, nous avons imposé des contraintes spécifiques qui ont conduit à développer nos concepts de paysages interactifs tactiles et sonores.

Nous avons détaillé l'avancement des premières versions des composants qui constituent deux dispositifs de transposition de la sémantique morpho-dispositionnelle. Les principaux développements à court terme concernent l'optimisation des algorithmes de segmentation de pages Web et d'extraction de mots-clés, et l'amélioration de la séparation perceptive des sources sonores par des techniques binaurales.

Dans une approche énactive, et en considérant son constat qu'il n'y a pas de perception sans action, il s'agit également d'initier une boucle vertueuse en ajoutant des fonctionnalités interactives dans le système. Les travaux envisagés portent sur cette intégration pour manipuler le tonnerre de mots et permettre une navigation basée sur une alternance aisée entre accès global et local à la page Web ; jusqu'à la découverte de l'information textuelle désirée.

Enfin, nous devons construire des expériences pour évaluer avec des personnes aveugles le système Tagthunder de *skimming* sonore et de nouvelles interactions multimodales visant à combiner cette solution avec notre système de *scanning* basé sur le dispositif tactile Tactinet.

Le travail algorithmique le plus abouti sur lequel nous avons récemment porté notre effort concerne l'amélioration des deux dispositifs. Dans les deux cas il est nécessaire de construire un module de segmentation automatique de pages Web performant, qu'il s'agisse d'exploiter les zones extraites pour façonner nos paysages tactiles ou sonores. Le dernier chapitre est consacré spécifiquement à décrire nos principaux apports à cette problématique.

8. Résultats autour de la segmentation automatique de pages Web

Au cours des dernières années, la création de contenu est devenue plus sophistiquée, multimédia et modulaire [165]. En particulier, les systèmes de gestion de contenu permettent de placer différents blocs sur une page Web d'une manière qui semble cohérente pour l'utilisateur, servant ainsi des objectifs spécifiques (distinction des sujets, mise en évidence des fonctionnalités, informations publicitaires, etc.). La segmentation de pages Web vise à identifier automatiquement ces régions sémantiques cohérentes. Elle peut être définie comme le processus consistant à diviser une grande page Web en plus petites régions, dans lesquelles les contenus ayant une sémantique cohérente restent ensemble [23].

Ce pré-traitement a montré qu'il améliorait la précision des tâches d'exploration du Web telles que la détection des doublons [29, 77], l'expansion des requêtes [23] et l'indexation [3]. Il existe une variété d'autres applications qui bénéficient de la segmentation automatique de pages Web. Par exemple dans le cadre d'annotations d'images pour l'amélioration de la précision apportée lorsque les annotations sont alimentées par les paragraphes auxquels elles appartiennent [22]. En extraction d'information également, la segmentation de pages Web peut être utilisée pour l'identification de sections de documents à forte densité de données [164] ou des champs de données isolés [109]. Dans nos travaux, et dans une approche dirigée par la tâche, l'application est développée à l'endroit des personnes non-voyantes en intégrant des algorithmes de segmentation de pages Web adaptés aux objectifs et aux contraintes de l'architecture modulaire du projet *Tagthunder* (Figure 8.1).

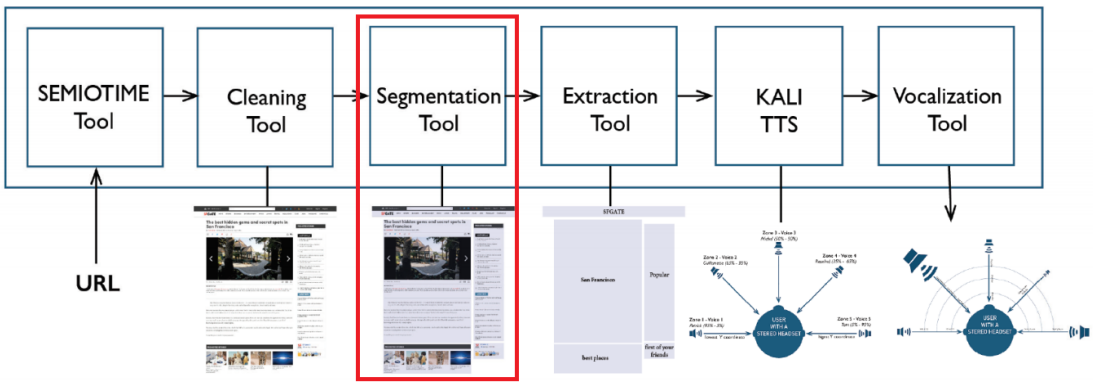


FIGURE 8.1. – Architecture modulaire : le module de segmentation

De manière générale, bien que la segmentation de pages Web semble être une tâche rela-

tivement aisée pour un humain, mesurer automatiquement la cohérence sémantique entre les différents contenus est une tâche difficile. En particulier lorsque le défi est de prendre en compte leur sémantique morpho-dispositionnelle [102, 168, 169].

Les modèles d'apprentissage supervisé qui s'y sont attelés se sont concentrés sur la délimitation du contenu et du bruit [159, 160], en se basant éventuellement sur les conclusions de [31], qui distingue très tôt cinq types de blocs : l'en-tête, le pied de page, la barre latérale gauche, la barre latérale droite et le contenu principal. D'autres tentatives se sont concentrées sur la possibilité d'exploiter la structure arborescente d'une page Web pour découvrir les frontières des zones [18, 29]. Cela dit, la supervisation nécessite évidemment un très grand ensemble de pages manuellement segmentées pour prendre en compte le large spectre de créativité de la conception Web. Là est certainement le principal goulot d'étranglement de ces méthodologies, et il est peu probable qu'elles puissent être utilisées dans des situations réelles en domaine ouvert.

Des stratégies non supervisées ont été proposées[23, 171, 29, 77, 3, 144, 175, 38, 8, 75, 76]. Dans ce contexte, les plus répandues sont basées sur des modèles *ad hoc*, qui reposent fortement sur des heuristiques définies manuellement et dépendent potentiellement de paramètres fixés expérimentalement [23, 171, 77, 144, 137, 175, 8, 75]. Pour éviter ces inconvénients, certaines approches s'appuient sur la théorie des graphes [29] ou utilisent des algorithmes classiques tels que le partitionnement agglomératif hiérarchique, ou en k-moyennes ou encore basés sur la densité [3, 8]. Ces solutions évitent la définition d'un ensemble cohérent d'heuristiques, limitent l'effort intensif d'essais et d'erreurs pour évaluer les multiples combinaisons possibles associées et sont plus susceptibles d'obtenir des optima globaux. Néanmoins, leur mise en oeuvre s'avère difficile pour être être portée en temps réel [50, 13, 149, 62, 143]. D'autres études se sont concentrées sur des techniques empruntées à la vision par ordinateur [38, 76] en adaptant les modèles de segmentation de photos ou de documents numérisés à la structure des pages Web. A notre connaissance, les résultats montrent une performance faible et une certaine lenteur d'exécution [76] confirmant nos propres expériences[6].

Dans ma recherche, l'intérêt pour cette problématique est apparu dès nos premiers travaux dans le cadre du projet *Tactinet* (*Cf.* chapitre 6). Depuis notre premier algorithme *Top Down Botton Up* (TDBU - [137]) en 2014, reposant sur une approche *ad hoc* basée sur les principes de la Gestalt, nous avons traversé toutes les questions sous-jacentes à cette introduction. Les sous-sections suivantes s'attachent à décrire nos évolutions dans ce domaine par des approches non supervisées. Elles s'appuient chronologiquement sur 3 articles clés qui décrivent nos principaux algorithmes de segmentation de pages Web et leur évaluation : [7] pour la section 8.1 ; [8] pour la section 8.2 ; et [73] pour la section 8.3.

8.1. Algorithmes KM, FKM et GE

Résolument tournés vers notre tâche d'amélioration de l'accessibilité non visuelle des pages Web, nous avons étudié dans un premier temps différentes stratégies de partitionnement en fixant des contraintes inhabituelles aux méthodes classiques. Nous avons considéré la segmentation de pages Web comme un problème de regroupement d'éléments visuels, où (1) **tous** les éléments

doivent être regroupés, (2) un nombre **donné** de regroupements doit être découvert, et (3) les éléments d'un regroupement doivent être visuellement **connectés**. Nous avons élaboré trois algorithmes différents qui respectent ces contraintes : en K-moyennes (KM), en F-K-moyennes (FKM) une variante de K-moyennes qui introduit la notion de force entre les éléments au lieu de la distance euclidienne, et un algorithme par expansion guidée ou *Guided Expansion* (GE) qui suit une stratégie de propagation incluant des contraintes d'alignement des partitions et de similarités visuelles. Nous avons mené une évaluation manuelle des trois algorithmes avec trois experts et deux mesures qualitatives : la compacité et la séparativité. Cependant, pour éviter les potentiels biais de subjectivité introduits par une évaluation humaine, nous avons proposé dans un second temps une évaluation quantitative à partir de différents critères d'analyse.

8.1.1. Travaux connexes

Dans la même période que nos premiers travaux, [144] a proposé *Block-O-Matic*, une stratégie qui décompose le processus de segmentation en trois phases d'analyse de contenu, de compréhension géométrique et de reconstruction logique de la page. Cette approche contrevient triplement aux contraintes algorithmiques que nous nous sommes fixées. D'abord elle s'appuie fortement sur la représentation interne et arborescente de la page (le *Domain Object Model* ou DOM), qui peut être fortement non alignée sur sa structure visuelle [175]. De plus, le nombre de partitions n'est pas déterminé en entrée du processus mais varie en fonction des décisions de l'algorithme. Enfin, une partition peut contenir des éléments isolés déconnectés des autres.

D'autres stratégies basées sur la structure visuelle ont été proposées. Les travaux notables qui suivent ce paradigme sont VIPS [23] et l'algorithme BCS (*Box Clustering Segmentation*) [175]. Alors que VIPS utilise toujours le DOM en combinaison avec des caractéristiques visuelles, BCS s'appuie exclusivement sur une représentation visuelle plate du document, ce qui permet une grande adaptabilité aux nouveaux contenus Web. En particulier, BCS suit un algorithme de regroupement agglomératif hiérarchique qui inclut un seuil, lequel contrôle le rassemblement des éléments visuels en groupes. En conséquence, le nombre de zones cohérentes est automatiquement déterminé par le seuil et donc soumis à variation, et certains éléments peuvent rester non groupés, de manière similaire à *Block-O-Matic* [144].

Notre solution s'appuie sur la même stratégie que l'algorithme BCS en nous basant exclusivement sur des éléments visuels pour orienter la segmentation, et donc sur une structure plate. En complément, nous proposons trois algorithmes de partitionnement différents qui respectent les contraintes imposées par notre tâche : (1) une segmentation en exactement **5 zones cohérentes**, (2) la **complétude** de la segmentation, où tous les éléments visuels appartiennent à exactement une partition et (3) la **connectivité** de tous les éléments à l'intérieur d'une partition donnée.

8.1.2. Stratégies de partitionnement évaluées

L'ensemble des évaluations présentées dans ce chapitre ont été menées sur un corpus de pages Web construit dès nos premiers travaux. Celui-ci est composé de 900 pages récupérées à partir de 300 noms de domaine répartis sur 100 sites de e-commerce, 100 sites de e-tourisme et

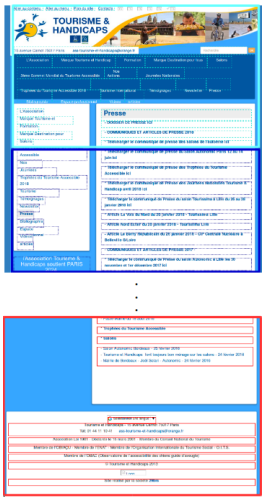


FIGURE 8.2. – KM

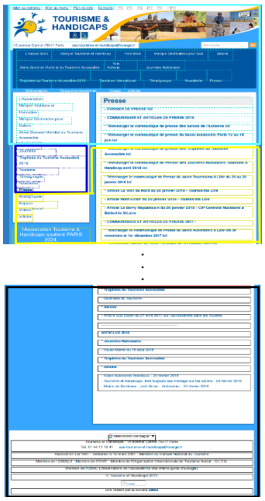


FIGURE 8.3. – FKM

FIGURE 8.4. – GE

FIGURE 8.5. – Man.

100 sites d'information. Pour chaque domaine nous avons extrait 3 pages Web (la page portail et 2 pages internes). Notre corpus final est ainsi constitué de ces 900 pages dont les noeuds des arbres DOM construits par le client Web ont été automatiquement annotés de 4 paramètres additionnels : (1) l'ensemble des styles CSS appliqués au noeud, conservant ainsi toutes les règles ayant un impact visuel sur l'élément HTML correspondant (*data-style*), (2) le XPATH de l'élément HTML permettant de conserver la position du noeud dans l'arbre DOM (*data-xpath*) ainsi que (3) les informations de position et de taille de la zone couverte physiquement par l'élément HTML (*data-bbox*), et (4) la visibilité effective de l'élément HTML qui peut parfois se trouver explicitement caché par sa position à l'extérieur de l'écran, par sa dimension, par une règle d'affichage particulière ou par l'action d'un programme dynamique côté client (*data-clean*).

Notre tâche spécifique de lecture non visuelle peut être définie à ce niveau du module de segmentation de pages Web comme un problème de partitionnement, pour lequel toutes les unités visuelles de base (derniers noeuds de style « block » dans notre arbre DOM enrichi) doivent être toutes rassemblées en un nombre fixe de 5 groupes cohérents. Afin d'explorer les solutions, nous avons proposé d'étudier trois stratégies algorithmiques différentes, *K-means* (KM), *F-K-means* (FKM) et *Guided Expansion* (GE), dont nous décrivons rapidement les principes généraux dans les sous-sections suivantes.

K-Means (KM)

Cette méthode classique [93] est favorable à la découverte d'un nombre de partitions fixées à l'avance. Dans notre contexte de segmentation automatique de pages Web, quelques adaptations ont malgré tout été nécessaires. Par exemple, la phase d'affectation d'un nouvel élément isolé à

un groupe en cours de construction est basée sur la plus courte distance euclidienne entre deux éléments. Pour notre tâche, dans laquelle les éléments sont des blocs rectangulaires, plusieurs alternatives étaient envisageables pour évaluer cette distance ; entre les centres des rectangles ou bien la plus courte entre deux bordures. C'est d'ailleurs cette dernière qui nous a paru la plus appropriée pour rendre compte d'une approche basée sur la structure visuelle des formes.

De plus, l'algorithme repose sur la sélection aléatoire de graines initiales dont la position peut avoir un impact considérable sur le résultat du partitionnement. Notre objectif étant de comparer différents algorithmes, nous proposons au contraire de fixer les 5 graines initiales en les répartissant de manière équidistante sur une diagonale de la page. Ce choix revêt un aspect arbitraire et donc discutable mais s'appuie dans un premier temps sur l'idée que la lecture rapide que nous essayons de modéliser se concentre sur des zones particulières de la page Web se situant sur la diagonale. D'autres stratégies seraient à étudier [54, 126]. Une illustration de l'utilisation de KM sur une page Web réelle est donnée par la figure 8.2.

F-K-Means (FKM)

Pour ce second algorithme, nous proposons une variante de KM qui prend plus fortement en compte la surface couverte par chaque élément visuel de base ; nous partons de l'idée que les éléments visuellement plus grands sont plus susceptibles « d'absorber » les plus petits que le contraire. Ainsi, si deux éléments visuels sont proches l'un de l'autre, leur fonction d'affectation $force(b_1, b_2)$ dépendra également de la différence des surfaces couvertes, telles que définies dans l'équation 8.1, où a_{b1} (resp. a_{b2}) est la surface de l'élément visuel ; $dist(.,.)$ représentent la plus courte distance euclidienne de bordure à bordure entre les éléments visuels.

$$force(b_1, b_2) = \frac{(a_{b1} * a_{b2})}{dist(b_1, b_2)} \tag{8.1}$$

Une illustration du FKM sur une page Web réelle est donnée dans la figure 8.3.

Guided Expansion (GE)

Avec l'algorithme *Guided Expansion* un seul élément visuel à la fois est assigné à un centroïde (point central d'un groupe candidat en construction), selon un ensemble de conditions qui incluent non seulement la plus courte distance euclidienne, mais également la similarité visuelle des éléments et l'alignement de leurs bordures.

$$vsim(\overrightarrow{b_1}, \overrightarrow{b_2}) = \sum_{i=1}^{|\overrightarrow{b_1}|} \mathbb{1}_{\overrightarrow{b_1^i} = \overrightarrow{b_2^i}} \tag{8.2}$$

Par exemple, la similarité visuelle $vsim(.,.)$ entre deux éléments b_1 et b_2 est calculée selon l'équation 8.2 à partir de leur vecteur de caractéristiques $\overrightarrow{b_1}$ et $\overrightarrow{b_2}$ composé par les propriétés

de style de la couleur d'arrière-plan de l'élément et de la police de caractère du texte qui y est inclus (famille, couleur, graisse).

$$dist(b, c) = argmin_{b_i \in c} dist(b, b_i) \tag{8.3}$$

$$vsim(\overrightarrow{b}, c) = argmax_{b_i \in c} vsim(\overrightarrow{b}, \overrightarrow{b_i}) \tag{8.4}$$

Notons que les métriques utilisées minimisent (pour la distance - 8.3) ou maximisent (pour la similarité visuelle - 8.4) la comparaison entre l'élément (b) et l'ensemble des éléments (b_i) de la partition en cours de construction (c).

Une illustration de GE sur une page Web réelle est donnée dans la figure 8.4.

8.1.3. Évaluation

Des évaluations qualitatives du partitionnement peuvent être menées. Elles s'appuient sur des experts afin de valider les résultats proposés par les algorithmes en les comparant à une vérité de terrain humaine [24]. Afin de limiter les biais potentiels de subjectivité, des études proposent des évaluations quantitatives reposant sur des mesures de corrélation des partitions. Dans ce contexte, [175] compare par exemple les algorithmes BCS et VIPS (voir sous-section 8.1.1) en utilisant les métriques usuellement exploitées dans ce domaine de recherche, le *F-score* et *l'indice de Rand ajusté*. Cependant notre problématique spécifique s'accommode mal de ces techniques d'évaluation classique de partitionnement. Par exemple, un seul élément incorrect peut briser en profondeur la structure visuelle, logique ou thématique d'une page Web tout en ayant un impact minime sur la métrique d'un point de vue quantitatif. De même, [145] crée une base de données de vérité terrain en segmentant des pages Web à l'aide de l'outil de partitionnement MoB [131] ; les auteurs proposent à partir de là des métriques spécifiquement adaptées, mais limitées à des méthodologies basées sur le DOM ; duquel nous ne voulons pas dépendre de manière trop stricte en raison de son fréquent non alignement sur la structure visuelle de la page Web qu'il participe à calculer. Pour ces raisons notre approche fut d'abord qualitative.

Évaluation qualitative

Notre première démarche a été de consulter 3 experts pour qu'ils produisent leurs propres segmentations de 53 pages Web extraites de notre corpus (constituant 3 vérités terrain individuelles - illustration donnée dans la figure 8.5). Pour apprécier et comparer ce corpus construit manuellement avec les résultats obtenus automatiquement, nous avons demandé aux experts de s'appuyer sur deux indices souvent utilisés pour l'évaluation d'algorithmes de partitionnement, les mesures de compacité et de séparation [1].

Projeté sur notre problématique, la **compacité d'une partition** mesure (entre 0 et 4) à quel point les éléments qui la composent se retrouvent également regroupés dans une même partition de la vérité terrain.La **compacité d'une page Web partitionnée** est la moyenne des compacités des partitions.

La **séparation d'une page Web partitionnée** mesure quant à elle la force des différences de découpage selon qu'il a été opéré par un algorithme ou par l'expert. Ce dernier devait évaluer en moyenne, sur une échelle également à 5 niveaux, dans quelle mesure les éléments d'une même partition de la vérité terrain étaient répartis dans des partitions différentes par un algorithme donné.

$$GS = \frac{(1 + separat) \times (1 + \overline{compact})}{|\text{evaluation scale}|^2} \tag{8.5}$$

Un score global (GS) était ainsi calculé par l'équation 8.5 (l'échelle du dénominateur fait référence à celle des notations, soit 5 dans cette évaluation).

Les experts devaient effectuer les segmentations manuelles puis affecter les notes de compacité et de séparation aux 3 partitionnements des 53 pages Web sans connaissance des algorithmes évalués.

Dans toutes les conditions testées, les résultats de cette évaluation qualitative montrent une **supériorité statistique de l'algorithme GE** aussi bien en terme de compacité que de séparation, et ce pour les trois experts ; ainsi qu'une **supériorité de KM sur FKM** (pour 2 experts). Cependant, il est à noter pour les trois algorithmes une évaluation en moyenne plus faible pour leur capacité de séparation que pour la compacité des partitions produites ; autrement dit, découvrir automatiquement des zones cohérentes entre elles semble une tâche plus difficile que construire des zones ayant une sémantique de construction interne cohérente.

De plus si les deux versions de l'algorithme KM et FKM sont comparables en terme de compacité, FKM obtient les résultats les plus faibles en matière de séparation. Nous observons une nette tendance de cette solution à produire des partitions fortement déséquilibrées en surface avec des écarts-types de mesures de compacité plus élevés : la tendance de FKM à créer des partitions très compactes mais petites ou des grandes partitions aérées pénalise fortement notre mesure de séparation.

Finalement, ces observations sont statistiquement confirmées par un test non-paramétrique des rangs signés de Wilcoxon effectué sur le score global GS ; l'observation **FKM < KM < GE** semble validée.

Évaluation quantitative

Si la force des résultats de l'évaluation qualitative peut être minimiser par les potentiels biais de subjectivité des avis experts, les discussions et analyses issues du processus nous ont permis de faire émerger des critères intéressants pour un jugement objectif du partitionnement. De là nous avons proposé 3 métriques exploitables pour associer automatiquement un score au partitionnement d'une page Web par un algorithme. Les critères numériques sur lesquels nos métriques s'appuient sont : (1) le nombre de contraintes logiques cassées ; (2) l'équilibre des partitions en terme de taille de la surface couverte, du nombre d'éléments visibles intégrés et de la quantité de texte qu'ils contiennent ; et (3) le chevauchement géométrique potentiel des rectangles théoriques englobant chaque partition (rectangles extérieurs).

(1) Bien Que nous avons évoqué plusieurs fois le risque de s'appuyer trop fortement sur l'arbre DOM participant au calcul par le client Web de la position des éléments qui composent la page, certaines contraintes de sa logique arborescente semblent malgré tout plus difficiles à violer. Ce sont ces ruptures fortement pénalisées par nos experts que que notre première métrique dénombre : ruptures aux frontières internes des items d'une énumération ordonnée ou non, ruptures entre un titre et le paragraphe qui le suit, ou encore au sein d'une entête, d'un pied de page ou d'un bloc de navigation.

(2) Il est apparu lors de l'évaluation qualitative que pour être correctement évaluée, une page Web devait avoir une composition équilibrée, c'est à dire être structurée en partitions ni trop différentes, ni trop similaires sur plusieurs paramètres formels. Aussi, chaque partitionnement de page Web reçoit un score fonction de l'écart-type entre toutes les partitions pour 3 critères d'équilibre (surface, éléments visibles, quantité de textes).

(3) Les experts ont également eu tendance à évaluer négativement des partitions entrelacées les unes aux autres. Pour évaluer ce phénomène, nous avons calculé le nombre de chevauchements entre les rectangles extérieurs de toutes les partitions (plus petit rectangle incluant tous les éléments d'une partition donnée).

	Ruptures Avg. $\pm\sigma$	Surfaces Avg. $\pm\sigma$	Textes Avg. $\pm\sigma$	Éléments Avg. $\pm\sigma$	Rect. ext. Avg. $\pm\sigma$
KM	2.12 $\pm$2.05	11.80 $\pm$6.46	11.40 $\pm$5.52	10.95 $\pm$8.01	5.21 $\pm$2.54
FKM	2.80 $\pm$2.76	21.14 $\pm$8.18	18.55 $\pm$7.74	22.79 $\pm$16.73	4.54 $\pm$2.20
GE	1.47 $\pm$1.85	17.34 $\pm$6.95	16.78 $\pm$6.37	19.67 $\pm$13.47	5.39 $\pm$2.22

TABLE 8.1. – Critères et résultats pour l'évaluation automatique de KM, FKM et GE de 150 pages Web.

Les critères numériques ont été calculés en terme de valeur moyenne et d'écart-type pour un ensemble de 150 pages extraites de notre corpus et partitionnées avec les trois algorithmes KM, FKM et GE. Les résultats sont regroupés dans la table 8.1.

Comme pour l'évaluation quantitative, les résultats montrent une supériorité significative de GE sur KM pour minimiser le nombre de ruptures ; ainsi que de KM sur FKM : $FKM_{rupt.} < KM_{rupt.} < GE_{rupt.}$.

Le critère d'équilibre semble conduire aux mêmes résultats quels que soient les 3 paramètres observés (surfaces, éléments, textes). A l'instar de l'évaluation quantitative, le FKM produit le plus grand déséquilibre et KM le plus petit. Le GE semble avoir un comportement plus proche de l'expert humain avec une tendance à générer uniquement de légers déséquilibres. En terme de significativité statistique $FKM_{eq.} < KM_{eq.}$ et $FKM_{eq.} < GE_{eq.}$

Le troisième critère, en calculant le nombre de chevauchements des « rectangles extérieurs » englobant les partitions, visait finalement à évaluer le nombre de partitions non rectangulaires. Le comportement semble similaire pour tous les algorithmes avec environ 5 (+ ou - 2) chevauchements en moyenne pour les 150 pages. Il y a cependant une tendance statistique pour FKM à produire moins de chevauchements, probablement en raison de sa propension au déséquilibre ; combiner une grande zone avec plusieurs petites rend peu probables leur chevauchement et donc la création de partitions non rectangulaires. Il serait intéressant d'approfondir et d'affiner

ce critère car nous observons une forme de paradoxe en comparant segmentations manuelles et automatiques : d'un côté les experts ne produisent quasiment aucune proposition intégrant des chevauchements alors que les algorithmes sont moins stricts sur ce critère ; d'un autre côté cette solution semble parfois tout à fait satisfaisante (en atteste la figure 8.4 qui semble cohérente en rassemblant les menus de cette manière). Une observation empirique plus approfondie nous a d'ailleurs permis de constater que l'algorithme GE semble plus performant également sur ce critère en produisant de meilleures zones non rectangulaires. Néanmoins, une étude plus poussée devra porter sur la manière de discriminer automatiquement les « bons » des « mauvais » chevauchements.

Bilan intermédiaire

En abordant la segmentation automatique de pages Web comme un problème de regroupement dirigé par la tâche de *skimming* non visuel, nous avons adapté l'algorithme classique KM et conçu deux autres algorithmes, FKM et GE, respectant les contraintes d'un nombre fixe de zones, de la complétude de la couverture et la connectivité des éléments visuels. Nous avons montré que les évaluations humaines et automatiques permettent de classer les algorithmes selon leur efficacité en fonction de plusieurs paramètres (le nombre de ruptures internes à certains éléments HTML, le nombre de chevauchements entre les partitions et leur équilibre formel), jouant des rôle complémentaires spécifiques pour les mesures de compacité et de séparation.

D'après les évaluations qualitatives et quantitatives, l'algorithme GE semble être la solution la plus efficiente sur tous les critères. Sa supériorité est probablement en grande partie liée à l'introduction de la contrainte d'alignement dans le processus d'expansion. En effet, l'affectation d'un élément isolé à une partition en cours de construction à chaque étape de cet algorithme permet de prendre des décisions plus fines que les méthodes plus globales soutenant KM et FKM pour lesquels la contrainte d'alignement est plus difficile à intégrer convenablement.

Néanmoins, il existe des limites claires. En particulier, le processus de partitionnement est très sensible aux positions initiales des graines. En suivant une stratégie de lecture diagonale, nous avons remarqué une tendance des algorithmes à privilégier l'horizontalité lors du partitionnement, et ainsi à rendre difficile l'identification de partitions axées naturellement sur la verticalité. Un autre problème du même ordre pourrait expliquer les limites démontrées pour FKM. Cet algorithme est désavantagé par le fait qu'un départ malencontreux dans un élément de petite taille limitera fortement les possibilités d'extension de la partition ; la métrique $force(.,.)$ tendant à favoriser les déséquilibres de surface.

La suite de ces travaux s'est ainsi portée sur la recherche de stratégies de lecture optimales pour tous les algorithmes en étudiant l'effet de la position des graines initiales.

8.2. Algorithmes de partitionnement et positionnement des graines initiales

Dans une perspective de recherche orientée par la tâche de lecture rapide de pages Web, nous avons amélioré les premiers résultats décrits dans la section précédente en évaluant différentes stratégies d'optimisation du positionnement des graines initiales des 3 algorithmes KM, FKM et GE, paramètre sensible des calculs. A découlé de ce travail une méthode de pré-partitionnement pour l'algorithme GE qui a permis d'accélérer le processus et de réduire les déséquilibres de taille entre les partitions découvertes.

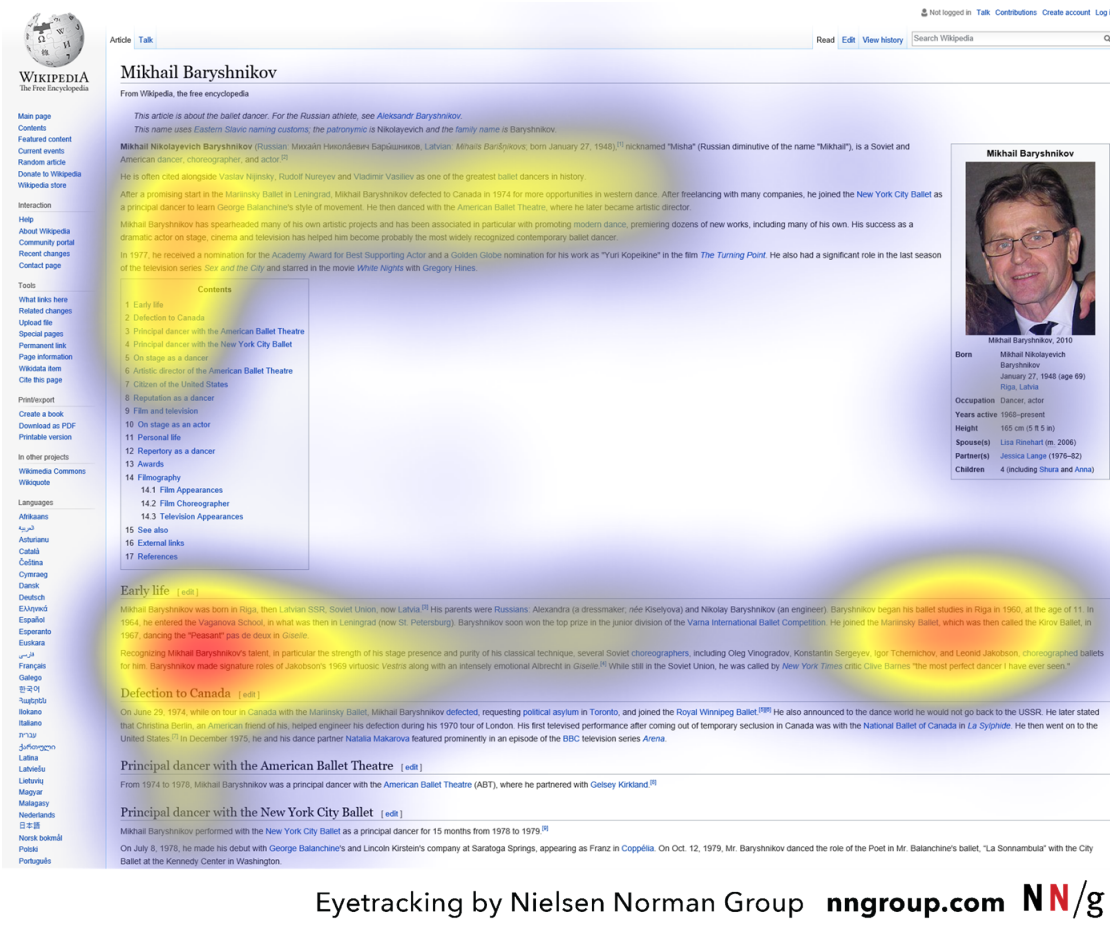


FIGURE 8.6. – Lecture en F d'une page Web extraite de [125]

[124] propose une étude sur la stratégie de lecture en « F » souvent observée (bien que parfois mal comprise) lors de la lecture de pages Web, même sur dispositifs mobiles : (1) les utilisateurs lisent d'abord dans un mouvement horizontal, généralement dans la partie supérieure de la zone de contenu. (2) Les utilisateurs descendent ensuite un peu dans la page et lisent dans un second mouvement horizontal qui couvre généralement une zone plus courte que le

mouvement précédent. (3) Enfin, les utilisateurs balayent le côté gauche du contenu dans un mouvement vertical. Les auteurs montrent notamment des cartes de fréquentation (*heatmaps*), qui mettent en évidence ce modèle (figure 8.6).

Une autre stratégie est étudiée par [11]. Les auteurs proposent une étude indiquant une lecture en « Z » lorsque les pages ne sont pas centrées sur leur contenu textuel : (1) les utilisateurs balaient d'abord du haut à gauche vers le haut à droite, formant une ligne horizontale ; (2) ensuite, vers le bas et le côté gauche de la page, créant une ligne diagonale ; (3) enfin, de nouveau vers la droite, formant une deuxième ligne horizontale.

Il est à noter que [11, 124, 125] suggèrent également d'autres stratégies utilisées sur le Web en fonction du type de contenus. Nos travaux se sont appuyés sur la lecture en « F » et en « Z » données comme les plus fréquentes.

8.2.1. Sélection des graines et pré-partitionnement de GE

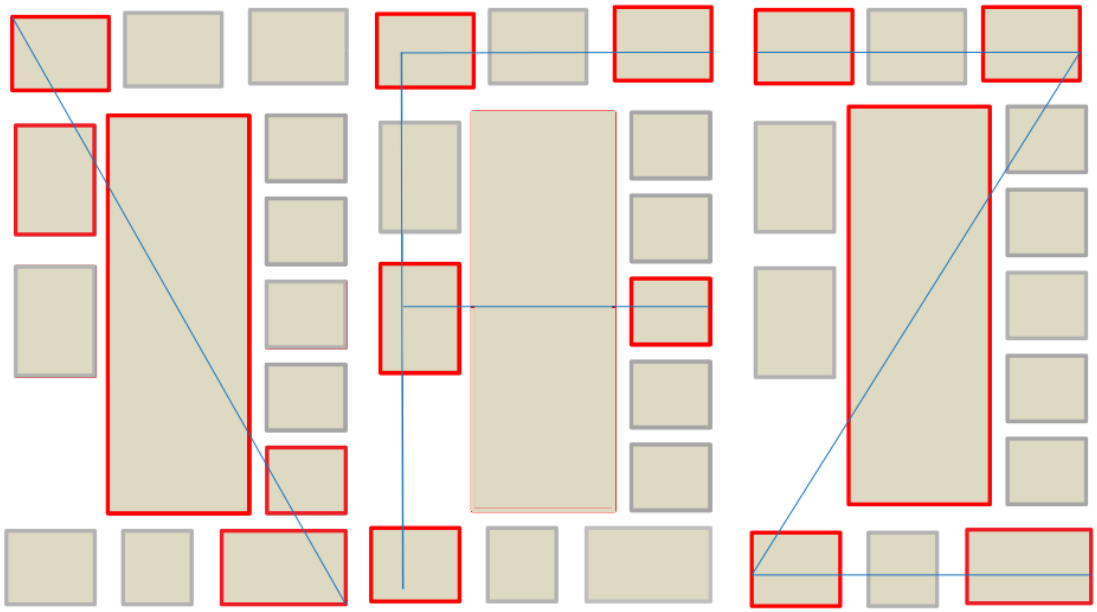


FIGURE 8.7. – Stratégies diagonale (gauche), F (centre) et Z (droite) pour positionner les graines initiales

La projection des stratégies de lecture étudiées sur le choix du positionnement des 5 graines initiales imposées par nos contraintes est indiquée par la figure 8.7. Les blocs représentent les éléments de base considérés pour une page Web donnée, les lignes bleues traversant les blocs représentent les stratégies de lecture et les blocs rouges indiquent les graines initiales choisies. Toutes choses égales par ailleurs avec l'étude précédente, nous avons conduit une évaluation quantitative des mêmes algorithmes KM, FKM et GE sur les mêmes critères de rupture, d'équilibre et de géométrie des partitions ; seules étaient modifiées les positions des graines initiales

et une optimisation de l'algorithme GE par un pré-partitionnement P afin de diminuer son ordre de complexité. Cette dernière stratégie préparatoire s'appuie sur l'algorithme QT (*Quality Threshold*) regroupant les éléments selon une zone de confiance déterminée par un seuil de distance [167]. Comme le suggère [175], cela peut être considéré comme un partitionnement à gros grain qui rassemble tous les éléments visuels de manière fiable autour d'une petite zone. Ce principe permet de déterminer un pré-partitionnement en 5 zones mais laisse des éléments non partitionnés. L'algorithme GE prend le relais pour finaliser le processus.

Enfin nous avons rajouté l'évaluation d'une variation de l'algorithme GE appelé FGE. Ce dernier prend en compte la surface couverte par chaque éléments visuels afin d'appliquer un premier critère de rapprochement d'une partition par le calcul d'une force d'attraction similaire à celle utilisée pour FKM. S'ensuit le processus original du GE qui s'appuie sur les paramètres d'alignement et de similarité visuelle.

8.2.2. Évaluation quantitative

Algorithm	Ruptures Avg. $\pm\sigma$	Surfaces Avg. $\pm\sigma$	Textes Avg. $\pm\sigma$	Éléments Avg. $\pm\sigma$	Rect. Ext. Avg. $\pm\sigma$
KM-D	2.12 $\pm$2.05	11.80 $\pm$6.46	11.40 $\pm$5.52	10.95 $\pm$8.01	5.21 $\pm$2.54
KM-F	2.59 $\pm$2.50	12.57 $\pm$6.54	12.52 $\pm$5.64	12.85 $\pm$9.63	4.13 $\pm$2.29
KM-Z	2.50 $\pm$2.40	13.20 $\pm$6.14	13.46 $\pm$6.02	14.85 $\pm$10.45	4.04 $\pm$2.21
FKM-D	2.80 $\pm$2.76	21.14 $\pm$8.18	18.55 $\pm$7.74	22.79 $\pm$16.73	4.54 $\pm$2.20
FKM-F	2.66 $\pm$2.40	20.58 $\pm$8.61	19.18 $\pm$8.63	23.87 $\pm$18.12	3.54 $\pm$1.94
FKM-Z	2.63 $\pm$2.36	21.14 $\pm$7.82	19.40 $\pm$7.57	25.32 $\pm$18.33	3.53 $\pm$1.95
GE-D	1.47 $\pm$1.85	17.34 $\pm$6.95	16.78 $\pm$6.37	19.67 $\pm$13.47	5.39 $\pm$2.22
GE-F	1.43 $\pm$1.85	22.64 $\pm$7.23	22.37 $\pm$6.70	30.42 $\pm$19.93	4.91 $\pm$2.01
GE-Z	1.34 $\pm$1.66	23.69 $\pm$7.10	22.77 $\pm$6.70	32.45 $\pm$21.82	5.26 $\pm$2.03
GE-P	1.57 $\pm$1.98	12.55 $\pm$6.76	12.24 $\pm$6.35	15.04 $\pm$11.12	6.72 $\pm$2.11
FGE-D	1.75 $\pm$1.94	28.50 $\pm$8.27	27.41 $\pm$7.74	38.80 $\pm$24.62	3.46 $\pm$1.89
FGE-F	1.83 $\pm$2.08	31.12 $\pm$7.29	29.65 $\pm$7.52	43.85 $\pm$25.21	3.53 $\pm$1.89
FGE-Z	1.77 $\pm$1.97	31.35 $\pm$6.88	30.26 $\pm$6.75	44.90 $\pm$25.96	4.18 $\pm$2.12
FGE-P	1.80 $\pm$2.15	13.70 $\pm$7.12	12.12 $\pm$6.70	14.64 $\pm$11.07	5.92 $\pm$2.36

TABLE 8.2. – Résultats de l'évaluation automatique de KM, FKM et GE pour toutes les stratégies de lecture D, F et Z. S'ajoute pour GE, la stratégie P avec pré-partitionnement. L'évaluation est effectuée sur 150 pages Web. $\pm\sigma$ représente la valeur de l'écart type sur la totalité des 150 pages Web.

Les 14 lignes du tableau 8.2 reprennent les 3 résultats du tableau 8.1 avec positionnement diagonaux (D) augmentés du résultat pour FGE avec la même stratégie et des 8 résultats obtenus pour toutes les combinaisons des algorithmes KM, FKM, GE et FGE avec les 2 stratégies de lecture en F et en Z. GE et FGE interrogent également les 2 combinaisons supplémentaires avec pré-partitionnement (P). Chaque colonne du tableau est également décrite sous la forme d'une boîte à moustache dans les figures 8.8, 8.9, 8.10, 8.11 et 8.12.

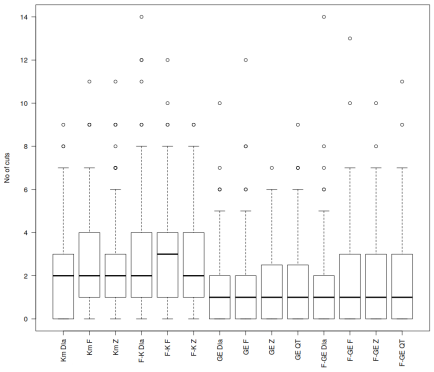


FIGURE 8.8. – Ruptures.

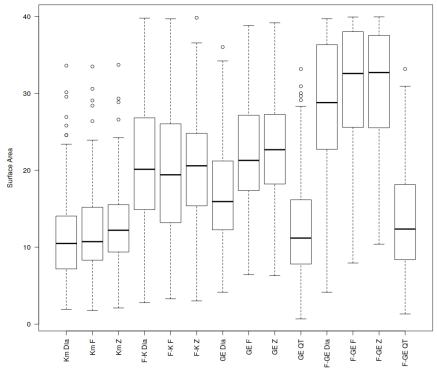


FIGURE 8.9. – Surfaces.

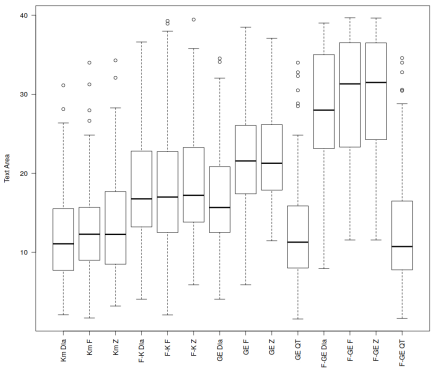


FIGURE 8.10. – Textes.

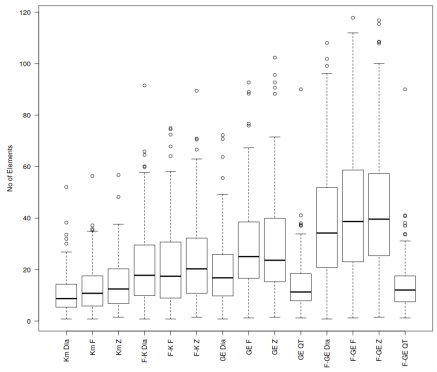


FIGURE 8.11. – Éléments.

Tout d'abord, les résultats montrent une nouvelle fois la supériorité de l'algorithme GE sur les autres algorithmes pour minimiser les ruptures ; et plus globalement la tendance est $FKM_{rupt.} < KM_{rupt.} < FGE_{rupt.} < GE_{rupt.}$. Les stratégies de lecture ne semblent pas jouer un grand rôle sur ce critère (figure 8.8).

Concernant le critère d'équilibre, dans tous les cas FGE montre les plus grosses variations entre partitions, tandis qu'à l'opposé, c'est KM qui présente les plus faibles. Comme dans l'expérience précédente, GE semble le plus proche de la segmentation manuelle par des experts avec son déséquilibre marqué mais limité ; sauf pour la version avec un pré-partitionnement qui semble pousser fortement vers un rééquilibrage des partitions (figure 8.9, 8.10 et 8.11).

Enfin, le critère mesurant le nombre de partitions non rectangulaires est comparable à la première expérience. D'après la figure 8.12, on peut remarquer que plus les graines sont placées près du bord de la page Web, plus elles ont tendance à faire des zones rectangulaires. En effet, la stratégie F place 5 graines près des bordures de la page Web et dénombre moins de chevauchements donc des zones plus rectangulaires ; la stratégie Z place 4 graines près des bordures et la

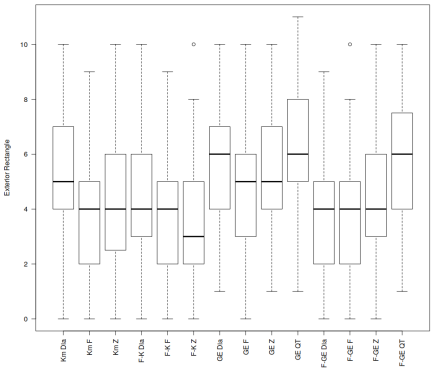


FIGURE 8.12. – Rectangles extérieurs.

stratégie diagonale en place 2 avec une corrélation dans le compte des chevauchements. Le pré-partitionnement semble être pourvoyeur du plus grand nombre de chevauchements. L'ordre des stratégies avec globalement moins de zones imbriquées peut être résumé par $P < D < Z < F$.

Enfin, pour déterminer avec précision quelles moyennes spécifiques sont significatives par rapport aux autres, nous avons conduit le test non paramétrique post hoc de comparaison multiple de Dunn. Les résultats de cette analyse sont présentés dans le tableau 8.3. Notez que les algorithmes au sein de chaque groupe ne sont pas significativement différents les uns des autres. Cependant, les algorithmes de différents groupes sont significativement différents les uns des autres. De plus, le rang de chaque groupe montre la performance d'un groupe donné pour un critère donné. D'après cette analyse, il semble que l'algorithme GE avec pré-partitionnement (GE-P) soit globalement la solution la plus adaptée à la segmentation de pages Web dans le contexte spécifique de l'accès à l'information non visuelle. Cet algorithme répond bien aux critères de jugement des experts en terme de minimisation des ruptures et de génération d'un équilibre marqué mais limité. Cependant, il est aussi celui qui produit le plus de chevauchements et, comme dans la première expérience, nous n'avons pas de critères définitifs pour faire la part des choses entre ceux qui seraient opportuns et les autres.

Au delà de cette dernière remarque, les résultats compilés dans le tableau 8.3 construisent un premier outil qui permettra d'adapter le choix d'algorithmes de segmentation en fonction des caractéristiques des pages Web envisagées, du domaine de la tâche visée et des critères qu'il s'agit de maximiser. Une difficulté qui émerge rapidement dans cette perspective est la complexité combinatoire mêlant des contraintes d'ordre visuel, logique ou sémantique ; et les compromis nécessaires et fragiles qu'elle sous-tend. C'est dans ce sens que nous avons conduit nos derniers travaux tournés vers des méthodes d'apprentissage automatique non supervisés adaptées à la recherche du meilleur équilibre entre tous ces objectifs à la fois. Ceux-ci sont présentés dans la dernière section de ce chapitre.

Critères	Groupes	
Ruptures	1	{GE-F, GE-Z, GE-P}
	2	{GE D}
	3	{FGE-D, FGE-P, FGE-Z}
	4	{FGE-F}
	5	{KM-D}
	6	{KM-F, FKM-D, FKM-Z}
	7	{FKM-F}
	8	{KM-Z}
Surfaces	1	{KM-D}
	2	{GE-P, KM-F, KM-Z}
	3	{FGE-P}
	4	{GE-D}
	5	{FKM-F}
	6	{FKM-Z}
	7	{FKM-D}
	8	{GE-F}
	9	{GE-Z}
	10	{FGE-D, FGE-F, FGE-Z}
Textes	1	{GE-P, FGE-P, KM-D, KM-F, KM-Z}
	2	{GE-D, FKM-D, FKM-F, FKM-Z}
	3	{GE-F, GE-Z}
	4	{FGE-D, FGE-F, FGE-Z}
Éléments	1	{KM-D}
	2	{KM-F}
	3	{GE-P, FGE-P, KM-Z}
	4	{GE-D, FKM-D, FKM-F, FKM-Z
	5	{GE-F, GE-Z}
	6	{FGE-D, FGE-F, FGE-Z}
Rect. Ext.	1	{FGE-D, KM-Z}
	2	{FGE-F}
	3	{FKM-F, FKM-Z}
	4	{KM-F}
	5	{FGE-Z}
	6	{FKM-D}
	7	{GE-F}
	8	{GE-Z, KM-D}
	9	{GE-D, FGE-P}
	10	{GE-P}

TABLE 8.3. – Test de Dunn pour les 14 algorithmes sur les 5 critères. Les algorithmes situés dans deux groupes différents présentent une différence statistique lors de leur comparaison sur un critère. Le rang montre l'ordre de performance pour chaque critère.

8.3. Algorithme MCS

Une approche différente de l'algorithme *K-means* a encore amélioré nos résultats en mettant en évidence le rôle clé de la sémantique morpho-dispositionnelle des page Web. Nous avons conçu une méthode de partitionnement multi-objectif (appelée MCS) dans laquelle (1) les indices visuels, logiques et textuels sont tous combinés de manière précoce et concomitante et (2) un processus évolutif découvre automatiquement le nombre optimal de partitions ainsi que l'initialisation correcte du processus. En tant que telle, notre proposition est sans paramètre, combine de nombreuses modalités différentes, ne dépend pas d'heuristiques réglées manuellement et peut être exécutée sur n'importe quelle page Web sans aucune contrainte.

Une évaluation exhaustive sur deux tâches différentes, où (1) le nombre de partition doit être découvert ou (2) le nombre de partition est fixé par rapport à la tâche à accomplir, montre que MCS améliore considérablement les algorithmes les plus compétitifs et les plus récents. En particulier, les résultats mettent clairement en évidence l'impact des modalités visuelles et logiques sur la performance de la segmentation.

8.3.1. Principales contributions

Algorithms		[23]	[171]	[29]	[77]	[3]	[144]	[175]	[8]	[75]	[38]	MCS
Cues	Visual	X	X	X	-	X	X	X	X	X	X	X
	Text	-	-		X	X	-	-	-	X	-	X
	Logical	X	X	X	-	X	X	-	-	X	-	X
Method	TD vs. BU	TD	TD	-	-	-	TD	BU	BU	-	TD	-
	AH vs. TH	AH	AH	TH	AH	TH	AH	AH	AH	AH	AH	TH
	ON vs. OF	ON	ON	OF	ON	OF	ON	ON	ON	ON	OF	OF
	PD vs. PF	PD	PD	PF	PD	PD	PD	PD	PD	PD	PD	PF
Evaluation	Manual	X	-	-	-	-	-	-	X	-	NA	-
	#EVI	-	1	2	2	1	5	2	-	3	NA	8
	#IVI	-	-	-	3	1	1	-	5	-	NA	4
	#RW	-	1	1	5	-	3	1	10	2	NA	7
	ET	1	-	1	1	-	-	-	-	-	NA	-

TABLE 8.4. – Stratégies non supervisées de segmentation de pages Web par types d'indices, méthodes d'apprentissage et cadres d'évaluation. Notez que TD (resp. BU) signifie Top-Down (resp. Bottom-Up), AH (resp. TH) signifie Ad Hoc (resp. Theoretical), ON (resp. OF) signifie On-line (resp. Off-line) clustering, PD (resp. PF) signifie Parameter-dependent (resp. Parameter-free) methodology, #EVI pour le nombre d'indices externes valides utilisés dans le cadre de l'évaluation, #IVI pour le nombre d'indices de validation internes, #RW pour le nombre de travaux connexes testés, et ET pour le nombre de tâches externes testées. Notez que NA signifie « informations non disponibles ».

Nous proposons un cadre d'apprentissage automatique non supervisé et multi-objectif appelé MCS basé sur l'algorithme *K-means*. MCS trouve automatiquement le nombre optimal de

partitions et positionne correctement les graines initiales. À cette fin, une stratégie de fusion précoce combine toutes les modalités ; différente métriques de distance et quatre fonctions objectifs s'attachent à les optimiser simultanément. Enfin, la recherche d'un optimum de Pareto est effectuée pour sélectionner les meilleures solutions ; c'est à dire celles présentant une satisfaction maximale (aucune amélioration sur une modalité ne peut se faire sans dégradation d'une autre).

Les métriques de distance utilisées s'appuient sur des indicateurs visuels, logiques et de sémantique textuelle. Pour les indices visuels, deux indicateurs sont pris en compte : la distance de bordure à bordure et la distance d'alignement entre les éléments visuels de base. Pour les caractéristiques logiques, deux distances différentes sont proposées à partir de la structure de l'arbre DOM calculé par le client Web pour représenter une page donnée : la longueur des chemins entre deux noeuds d'une part et les préfixes communs des expressions *xpaths* de chaque élément d'autre part (chemins de localisation depuis la racine dans l'arbre DOM). Pour les indices textuels, une valeur sémantique est calculée sur la base de plongements documentaires [82].

La performance de MCS est évaluée d'une part sur une tâche classique de segmentation de pages Web en un nombre de partitions variable non défini en entrée du processus ; et d'autre part, sur une tâche spécifique contrainte à la découverte d'exactement 5 zones adaptées au *skimming* non-visuel. Afin d'avoir une vue d'ensemble complète des résultats obtenus, nous calculons huit indices de validation externe et quatre indices de validation interne. Nous comparons les résultats de performance et la signification statistique avec sept autres algorithmes alternatifs. En particulier, nous proposons également deux nouvelles adaptations de l'algorithme GE-P décrit dans la section précédente.

Sur deux corpus de référence de 51 pages Web (un pour chaque tâche) extraites de notre corpus, les résultats expérimentaux montrent que MCS surpasse significativement tous les travaux connexes pour la plupart des indices de validation externe ; démontrant les avantages d'optimiser la combinaison des modalités visuelles, logiques et textuelles. Les sous-sections suivantes résumes les principes de la démarche en terme de méthodologie et de résultats. Un état de l'art critique des méthodes non supervisées de segmentation de pages Web est structuré dans le tableau 8.4. Une description plus détaillée de cette analyse ainsi que les tenants et aboutissants de ce travail peuvent être consultés dans [73].

8.3.2. méthodologie globale

Nous proposons de nous appuyer sur les travaux récents de [142] dédiés au partitionnement multi-objectif basé sur l'algorithme K-means. Les techniques de partitionnement traditionnelles optimisent implicitement une seule fonction objectif interne qui peut mesurer la compacité, la séparation spatiale, la connectivité, la densité ou la symétrie entre les partitions. Pour la segmentation de pages Web les partitions doivent présenter des spécificités selon différents points de vue : visuel, textuel ou logique. L'application de techniques d'optimisation multi-objectifs qui maximisent/minimisent en même temps différents indices de validité des partitions est apparue comme une alternative prometteuse [67, 140, 142]. Pour cela, nous proposons un cadre qui combine un réseau de neurone artificiel par cartes auto-organisatrices de Kohonen avec un algorithme d'évolution différentielle multi-objectifs. Le caractère évolutif s'appuie sur un co-

dage des centres de partition (centroïdes) sous la forme de chromosomes maximisant au fur et à mesure du processus des objectifs concurrents et la qualité optimale du partitionnement.

L'originalité de MCS est de combiner un ensemble de représentations discrètes et continues en fonction du point de vue visuel, textuel ou logique. MCS peut ainsi être considéré comme un algorithme de partitionnement multi-critères, pour lequel différents objectifs sont optimisés simultanément par un algorithme génétique : différentes stratégies d'élagage et de croisement altèrent la reproduction artificielle des chromosomes afin d'explorer plus largement l'espace de recherche, réduire à chaque itération la distance de voisinage entre les chromosomes dans le réseau et ainsi améliorer la convergence. En revanche, aucune mutation n'est effectuée en raison des contraintes de représentation du problème.

Formulation du problème

Une page Web est interprétée comme un ensemble d'informations structurées par des éléments HTML, chacun d'entre eux possédant ses attributs propres, enrichis par nos soins avec l'ensemble des propriétés de style, de position, de surface et de visibilité qui s'y appliquent. Ces informations permettent en entrée de sélectionner les éléments candidats (surfaces rectangulaires représentant les éléments HTML visibles) mais également de participer aux calculs des fonctions objectifs. La tâche de segmentation de page Web peut être formalisée de la manière suivante.

- Donnée :
  - Une page Web avec $\mathbb{N}_e$ éléments $\mathbb{W} = e_1, e_2, ..., e_{\mathbb{N}_e}$, chacun possédant ses propres caractéristiques textuelles, visuelles et logiques.
  - Un ensemble de $\mathbb{N}_f$ fonctions objectifs $\mathbb{F} = F_0, F_1, ..., F_{\mathbb{N}_f}$, cahque F_i mesurant à quel point l'assignation d'un élément à une partition est optimisée.
  - Un intervalle $[Kmin..Kmax]$ pour encadrer les $\mathbb{N}_p$ partitions à découvrir, i.e., $Kmin \leq \mathbb{N}_p \leq Kmax$
- Résultat :
  - Une assignation $\mathbb{A} = A_0, A_1, ..., A_{\mathbb{N}_p}$ des $\mathbb{N}_e$ éléments tels que
    - $\forall A_i \in \mathbb{A}$, $A_i = \{e_1^i, e_2^i, ..., e_{T_i}^i\}, T_i = |A_i| > 0$ (pas de partition vide)
    - $\bigcup_{i=1}^{\mathbb{N}_p} A_i = \mathbb{W}$ and $\bigcap_{i=1}^{\mathbb{N}_p} A_i = \varnothing$ (partitionnement complet et sans recouvrements)
  - qui optimise simultanément toutes les fonctions objectifs dans $\mathbb{F}$, i.e. $\mathbb{A}$ appartient à un front optimal de Pareto.

Un chromosome encode un ensemble de centres de partitions, c'est-à-dire une assignation possible $\mathbb{A}$ correspondant à une solution de partitionnement. MCS tente de déterminer l'ensemble optimal qui partitionne une page Web en faisant varier la taille des chromosomes dans la plage $Kmin \leq \mathbb{N}_p \leq Kmax$. Pour générer la population initiale de la $i^{ème}$ solution, un

nombre K_i aléatoire de graines est disséminé aléatoirement parmi l'ensemble des éléments représentant la page Web. La méthode de partitionnement K-means enchaîne ensuite (1) les étapes d'assignation à partir de différentes métriques puis (2) les mises à jour.

Cette dernière étape de mise à jour à nécessité d'introduire le concept d'élément virtuel afin de favoriser la consistance des calculs en moyenne. Un élément atomique du partitionnement étant une surface rectangulaire contenant du texte, le calcul du centroïde d'une partition devrait être une surface rectangulaire moyenne qui n'existe pas réellement dans les données. Nous l'avons donc conceptualisée comme un élément virtuel v associé à deux paires de coordonnées et d'un texte. Initialement congruent à un élément réel de la page, l'élément virtuel, au grès des nouveaux éléments qui lui sont assignés, évolue vers la partition composant une solution finale. Chaque itération modifie les paramètres de v par (1) le calcul de la moyenne des coordonnées des éléments de la partition et (2) la concaténation de leur texte transformé dans un espace continu grâce au modèle de représentation numérique de documents *Doc2vec* [82].

Métriques de distance

Pour l'étape d'assignation qui vise à allouer de nouveaux éléments à un élément virtuel donné à chaque itération de l'algorithme, nous définissons trois familles différentes de distances (visuelle, logique et textuelle) comme autant de dissimilarités possibles entre éléments. Notez qu'une normalisation min-max est systématiquement appliquée à toutes ces distances.

Distance visuelle de bordure à bordure

$$
bbd_i^k = \begin{cases}
0 & \begin{cases} ((x_1' \leq x_1 \leq x_2') \vee (x_1' \leq x_2 \leq x_2') \vee (x_1 \leq x_1' \leq x_2' \leq x_2)) \wedge \\ ((y_1' \leq y_1 \leq y_2') \vee (y_1' \leq y_1 \leq y_2') \vee (y_1 \leq y_1' \leq y_2' \leq y_2)) \end{cases} \\
\sqrt{(x_1' - x_2)^2 + (y_1' - y_2)^2} & \begin{cases} (x_2 \leq x_1') \wedge (y_2 \leq y_1') \end{cases} \\
y_1' - y_2 & \begin{cases} ((x_1' \leq x_1 \leq x_2') \vee (x_1' \leq x_2 \leq x_2') \vee (x_1 \leq x_1' \leq x_2' \leq x_2)) \wedge \\ (y_2 \leq y_1') \end{cases} \\
\sqrt{(x_1 - x_2')^2 + (y_1' - y_2)^2} & \begin{cases} (x_2' \leq x_1) \wedge (y_2 \leq y_1') \end{cases} \\
x_1 - x_2' & \begin{cases} (x_2' \leq x_1) \wedge \\ ((y_1' \leq y_1 \leq y_2') \vee (y_1' \leq y_2 \leq y_2') \vee (y_1 \leq y_1' \leq y_2' \leq y_2)) \end{cases} \\
\sqrt{(x_1 - x_2')^2 + (y_1 - y_2')^2} & \begin{cases} (x_2' \leq x_1) \wedge (y_2' \leq y_1) \end{cases} \\
y_1 - y_2' & \begin{cases} ((x_1' \leq x_1 \leq x_2') \vee (x_1' \leq x_2 \leq x_2') \vee (x_1 \leq x_1' \leq x_2' \leq x_2)) \wedge \\ (y_2' \leq y_1) \end{cases} \\
\sqrt{(x_1' - x_2)^2 + (y_1 - y_2')^2} & \begin{cases} (x_2 \leq x_1') \wedge (y_2' \leq y_1) \end{cases} \\
x_1' - x_2 & \begin{cases} (x_2 \leq x_1') \wedge \\ ((y_1' \leq y_1 \leq y_2') \vee (y_1' \leq y_2 \leq y_2') \vee (y_1 \leq y_1' \leq y_2' \leq y_2)) \end{cases}
\end{cases}
\tag{8.6}
$$

Comme pour les approches décrites dans les sections précédentes de ce chapitre, nous utilisons la distance de bordure à bordure bbd_i^k entre un élément rectangulaires b_i et un élément virtuel v_k. Si (x_1, y_1) et (x_2, y_2) sont les coordonnées respectives haut-gauche et bas-droit de l'élément b_i, et (x_1', y_1') et (x_2', y_2') sont les coordonnées respectives haut-gauche et bas-droit de

l'élément virtuel v_k, bbd_i^k est calculé comme dans l'équation 8.6.

Distance visuelle d'alignement

$$ald_i^k = 1 - \max_{b_j \epsilon P_k^t, j \neq i} \left\{ \frac{is\_aligned(b_i, b_j)}{bbd_i^j + 1} \right\} \tag{8.7}$$

L'alignement, déjà exploité par les algorithmes de type GE, est un paramètre qualitatif que nous devons quantifier afin de l'adapter à une comparaison de distance entre éléments. Dans notre approche, deux éléments sont considérés comme alignés si leurs bordures horizontales ou verticales coïncident à plus ou moins 5 pixels près. Pour l'itération $t+1$, la distance d'alignement ald_i^k entre un élément b_i et un élément virtuel v_k représente l'élément géométriquement le plus proche et aligné avec l'élement virtuel. Elle est donnée par l'équation 8.7, où bbd_i^j est la distance de bordure à bordure entre les éléments b_i et b_j, $is\_aligned(b_i, b_j)$ est égal à 1 si b_i et b_j sont alignés et 0 sinon, et P_k^t est l'ensemble des éléments composant la partition P représentée par l'élément virtuel v_k à l'itération t.

Distance logique entre noeuds de l'arbre DOM

Il est fréquent que la structure logique de l'arbre DOM, représentation interne d'une page Web, soit partiellement éloignée de la structure logique et hiérarchique de l'information interprétée visuellement par le lecteur (en raison de l'application de règles de style modifiant en profondeur la visualisation de la page sans changements dans la structure de l'arbre DOM, ou bien de l'intervention dynamique et interactive de programmes informatiques directement sur les noeuds de l'arbre DOM). Cela rend difficile l'utilisation de cette représentation interne pour mesurer des distances entre structures logiques ; et conduit souvent à la laisser entièrement de côté dans les calculs. Notre approche multi-objectif nous permet de prendre indirectement en compte cette limitation et de réintégrer l'arbre DOM dans nos équations puisqu'il s'agira de rechercher un compromis optimal en minimisant les distances qui en sont issues sans dégrader les autres distances.

$$pathDist_i^j = l_i + l_j - 2l_{ij} + 1 \tag{8.8}$$

En s'inspirant de [75], nous utilisons les expressions *xpath* de deux éléments pour mesurer une distance entre les positions des noeuds de l'arbre DOM qu'elles représentent. Soit l_i (resp. l_j) la longueur de l'expression *xpath* de l'élément b_i (resp. b_j), et l_{ij} la longueur du préfixe commun, la distance $pathDist_i^j$ entre b_i et b_j est définie dans l'équation 8.8.

$$pd_i^k = \min_{b_j \epsilon P_k^t, j \neq i} \lceil pathDist_i^j \rceil \tag{8.9}$$

La distance pd_i^k entre un élément b_i et un élément virtuel v_k est définie à l'itération $t+1$ par l'équation 8.9, où P_k^t est l'ensemble des éléments assignés à l'élément virtuel v_k à l'itération t. Les distances normalisées entre deux noeuds sont notées $\lceil pathDist_i^j \rceil$.

Distance logique des expressions *xpath*

$$xpSim_i^j = \sum_{r=1}^{min(|\overrightarrow{b_i}|,|\overrightarrow{b_j}|)} \mathbb{1}_{\overrightarrow{b_i^r}=\overrightarrow{b_j^r}} \tag{8.10}$$

Les pages Web peuvent avoir des structures hiérarchiques répétées (page de produits commerciaux, d'articles de presse, etc.). Ces éléments composites, bien que dissemblables en termes de contenu, sont similaires en termes de structure des expressions *xpath* qui représentent leurs noeuds dans l'arbre DOM. Le score de similarité xpath $xpSim_i^j$ entre deux éléments b_i et b_j est défini dans l'équation 8.10, où $\overrightarrow{b_i}$ (resp. $\overrightarrow{b_j}$) sont les vecteurs encodant les expressions *xpath* de b_i et b_j, l'exposant r parcourant les noeuds traversées par les deux expressions *xpath*.

$$xpd_i^k = 1 - \max_{b_j \epsilon P_k^t, j \neq i} \lceil xpSim_i^j \rceil \tag{8.11}$$

La distance xpd_i^k entre un élément b_i et un élément virtuel v_k est définie à l'itération $t+1$ par l'équation 8.11, où P_k^t est l'ensemble des éléments assignés à l'élément virtuel v_k à l'itération t. Les distances normalisées entre deux expressions *xpath* sont notées $\lceil xpSim_i^j \rceil$.

Distance textuelle

Des éléments peuvent contenir des informations textuelles dont nous souhaitons mesurer la similarité sémantique en faisant l'hypothèse d'une cohérence lexicale dans chaque partition. En considérant les nombreuses méthodes récentes pour représenter les textes sous la forme de vecteurs de caractéristiques continues [82, 28, 108], nous proposons d'utiliser *Doc2vec* [82].

$$txd_i^k = \frac{1 - \frac{\overrightarrow{v_k^{txt}}.\overrightarrow{b_i^{txt}}}{||\overrightarrow{v_k^{txt}}||.||\overrightarrow{b_i^{txt}}||}}{2} \tag{8.12}$$

Le contenu textuel d'un élément virtuel v_k peut être défini comme la concaténation des contenus textuels de tous les éléments qui lui sont alloués et représenté par le vecteur numérique $\overrightarrow{v_k^{txt}}$. De la même manière, un vecteur continu de texte peut être obtenu pour chaque élément, noté $\overrightarrow{b_i^{txt}}$ pour l'élément b_i. La distance textuelle normalisée txd_i^k entre un élément virtuel v_k et un élément b_i peut être calculé sur la base de la mesure de similarité du cosinus, généralement utilisée pour la similarité textuelle. Cette distance est définie dans l'équation 8.12.

Combinaison de distances et fonction d'assignation

$$dist_i^k = \frac{1}{3}\left(\frac{(bbd_i^k + ald_i^k)}{2} + \frac{(pd_i^k + xpd_i^k)}{2} + txd_i^k\right) \tag{8.13}$$

La distance entre un élément virtuel v_k et un élément b_i est évaluée de différents points de vue visuels, logiques et textuels mais doit constituer une métrique unique pour exploiter l'algorithme

K-means. Cette distance $dist_i^k$ est définie dans l'équation 8.13.

$$m = \min_k dist_i^k \tag{8.14}$$

Sur cette base, la fonction d'assignation de l'algorithme K-means adapté peut facilement être définie avec l'équation 8.14.

Fonctions Objectifs

Nous définissons quatre fonctions objectifs qui exploitent les distances proposées dans la sous-section précédente. L'indice de Davies-Bouldin [43] est utilisé pour définir des objectifs géométriques et textuels, le coefficient de silhouette [134] permet de définir un objectif d'alignement, et une fonction objectif spécifique est définie pour évaluer la tendance d'une solution de partitionnement à produire des ruptures logiques non désirées, au sens déjà défini dans les travaux précédents.

Objectif géométrique *DB*-Bordure

L'indice de Davies-Bouldin (DB) est une mesure de compacité et de séparation d'une partition. Il nécessite la définition de deux fonctions pour mesurer le rapport entre la dispersion au sein d'une partition et la qualité de séparation entre les partitions. La performance du partitionnement est d'autant meilleure que l'indice est faible. Le processus évolutif vise alors à minimiser l'indice DB calculé à partir de ces fonctions de dispersion et de séparation.

$$S_k^{bb} = \left(\frac{1}{T_i} \sum_{i=1}^{T_k} (bbd_k^i)^2 \right)^{1/2} \tag{8.15}$$

Nous utilisons l'indice DB pour définir le premier objectif géométrique qui vise à favoriser la connectivité des éléments en s'appuyant sur notre mesure de la distance bordure à bordure. Nous définissons par l'équation 8.15 la fonction de dispersion S_k^{bb} pour la $k^{ième}$ partition, dont les éléments sont $b_1, b_2, ..., b_{T_k}$ et l'élément virtuel v_k.

$$M_{ij}^{bb} = M_{ji}^{bb} = bbd_i^j \tag{8.16}$$

La fonction de séparation M_{ij}^{bb} entre deux partitions P_i et P_j, est définie par l'équation 8.16 comme la distance bordure à bordure entre les éléments virtuels v_i et v_j.

Objectif textuel *DB*-Text

Nous utilisons l'indice DB pour définir l'objectif textuel qui vise à favoriser la cohérence interne des partitions et défavoriser leur cohérence externe en s'appuyant sur notre mesure de la distance textuelle.

$$S_k^{txt} = \left(\frac{1}{T_k} \sum_{i=1}^{T_k} (txd_k^i)^2 \right)^{1/2} \tag{8.17}$$

Nous définissons par l'équation 8.17 la fonction de dispersion S_k^{txt} pour la $k^{ième}$ partition, dont les vecteurs texte des éléments sont $b_1^{txt}, b_2^{txt}, ..., b_{T_k}^{txt}$ et le vecteur texte de son élément virtuel v_k^{txt}.

$$M_{ij}^{txt} = M_{ji}^{txt} = txd_i^j \tag{8.18}$$

La fonction de séparation M_{ij}^{txt} entre deux partitions P_i and P_j est définie par l'équation 8.18 comme la distance textuelle entre les vecteurs v_i^{txt} and v_j^{txt}.

Objectif d'alignement *SIA*

Il a été démontré depuis longtemps que l'alignement joue un rôle majeur dans la segmentation de pages Web [23, 171, 144, 8]. L'objectif d'alignement vise à favoriser les éléments alignés au sein des partitions, tout en défavorisant cet alignement entre les éléments de partitions différentes.

L'indice DB est une mesure basée sur des moyennes ; ce qui le rend peu adapté à une métrique par paire telle que l'alignement des éléments. Pour y remédier, nous proposons de nous appuyer sur le coefficient de silhouette [134]. Pour chaque élément b_i affectée à la partition P_p avec son élément virtuel v_p, et toute autre partition P_k avec son élément virtuel v_k, posons :

$$i_i = \sum_{j \in v_p, i \neq j} \mathbb{1}_{b_i \text{ is aligned } b_j}$$

$$e_i = \max_{k \neq p} \sum_{j \in v_k} \mathbb{1}_{b_i \text{ is aligned } b_j}$$

$$n_i = \sum_{j \in v_k, k \neq p} \mathbb{1}_{b_i \text{ is aligned } b_j} + i_i$$

$$sia_i = \frac{i_i - e_i}{\max\{n_i - i_i, n_i - e_i\}}$$

alors, le coefficient silhouette d'alignement SIA est défini par l'équation 8.19, où $\mathbb{N}_b$ est le nombre d'éléments dans la page Web.

$$SIA = \frac{1}{\mathbb{N}_b} \sum_{i=1}^{\mathbb{N}_b} sia_i \tag{8.19}$$

Objectif du nombre de ruptures logiques CUTS

Une proposition similaire aux fonctions objectifs précédentes mais basée sur les distances pd_i^k et xpd_i^k donnerait lieu à des indices erronés en raison de leurs définitions internes. Nous proposons que chaque fois qu'une contrainte logique est rompue, cela comptabilise une rupture ; chaque page Web est évaluée en fonction de son nombre total de ruptures. En conséquence, cette mesure sera calculée à partir des algorithmes déjà existants évoqués dans la section 8.1.3.

Sélection

Une fois que le K-means adapté a été exécuté sur les différents chromosomes de la population, et que chaque solution a été évaluée en fonction des quatre fonctions objectifs, l'étape suivante du processus évolutif consiste à sélectionner les meilleurs individus pour la reproduction. Cette étape combine un tri non dominé et un élagage de la population basé sur des cartes auto-organisatrices de Kohonen.

Tri non dominé

Chaque chromosome est associé à quatre valeurs d'objectif et le tri non dominé est utilisé pour sélectionner les meilleures solutions sur un ensemble de fronts optimaux de Pareto. Pour cela, à l'instar de [142], nous utilisons l'algorithme génétique de tri non dominé NSGA-II [47]. Afin de contrôler la taille de la population, si le nombre de chromosomes sélectionnés $|D|$ dépasse une $Soft\text{-}Limit$ prédéfinie, des cartes auto-organisatrices de Kohonen sont utilisées pour l'élagage de la population vers une $Hard\text{-}Limit$.

Cartes auto-organisatrices de Kohonen

Les cartes auto-organisées de Kohonen (SOM pour *Self-Organizing Maps*) [78] sont utilisées pour classer sans supervision l'ensemble des solutions afin de sélectionner les D meilleures. Les SOM sont constituées de nœuds u organisés dans une grille à deux dimensions, chaque nœud occupant des coordonnées cartésiennes fixes.

Dans le contexte de notre travail, chaque nœud est structurellement similaire à un chromosome ayant un ensemble de $Kmax$ éléments virtuels. Chaque chromosome de D doit être associé à un nœud. Deux chromosomes appartiennent à la même famille s'ils sont associés au même nœud. Un chromosome i est associé au nœud u' (dit nœud gagnant) si il minimise une distance nœud-chromosome $d(i, u)$ calculée à partir d'un algorithme glouton sur les distances bordure à bordure et textuelles. Enfin, les nœuds voisins u du nœud gagnant u' sont mis à jour pour les rapprocher.

Ce processus est répété avec un paramètre variable de taux d'apprentissage qui permet de limiter progressivement les mises à jour ; le but est d'assurer la stabilisation de la carte auto-organisatrice de Kohonen et donc la convergence au bout d'un nombre d'itération maximal donné.

Reproduction

Une fois que suffisamment de solutions sont sélectionnées pour la reproduction, l'opération d'enjambement (croisement) est effectuée pour obtenir une nouvelle population de chromosomes potentiellement plus performants.

Pour cela, une première solution est choisie au hasard dans la population et une seconde dans un voisinage d'autant plus proche que le nombre d'itération maximale est atteint (assurant ainsi la convergence du processus).

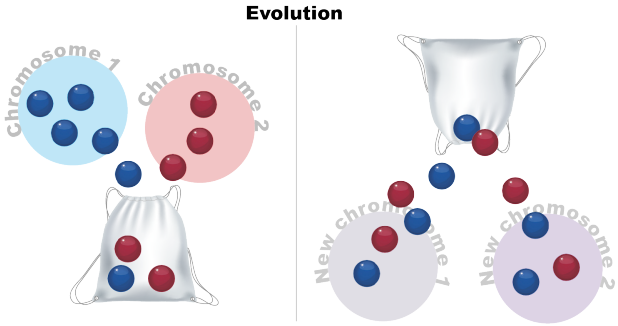


FIGURE 8.13. – Stratégie des sacs pour l'enjambement.

Tous les gènes individuels des deux chromosomes sont alors mélangés et reconstruits artificiellement selon la stratégie imagée dans la figure 8.13. Deux nouvelles solutions sont créées de sorte que chaque chromosome contienne entre $Kmin$ et $Kmax$ gènes dont au moins un de chaque chromosome initial.

Le processus de reproduction est ainsi réitéré jusqu'à épuisement des candidats.

Itération et terminaison

Une fois la nouvelle population créée, elle est fusionnée avec la population préexistante pour l'itération suivante du processus d'optimisation, de sorte que le tri non dominé est à nouveau exécuté et qu'un nouveau front optimal de Pareto est obtenu. Ce processus est itéré jusqu'à ce que le nombre d'itération maximal soit atteint.

Une solution optimale unique doit être alors extraite du dernier front. Cette question reste un problème ouvert comme mentionné dans [114]. Notre stratégie consiste en un de tri par priorité en sélectionnant la solution qui optimise un ordre spécifique des objectifs. Si nous notons A l'alignement du coefficient silhouette, T le DB-Text, G le DB-Border, et C le nombre de ruptures, toute combinaison de ces 4 lettres définira l'ordre de priorité décroissante des objectifs à optimiser pour sélectionner une solution. Par exemple, la combinaison d'objectifs ACGT indique de choisir la solution qui maximise l'alignement sur toutes les solutions ; en cas d'égalité, le deuxième objectif à minimiser sera le nombre de ruptures, et ainsi de suite, jusqu'à la

minimisation de la distance textuelle. Par souci d'exhaustivité, nous avons comparé les 24 combinaisons possibles afin de comprendre le poids de chaque critère (visuel, logique, textuel) pour la segmentation de pages Web. Ces résultats font partie des évaluations menées et résumées dans la dernière sous-section.

8.3.3. Évaluations quantitatives

Afin d'évaluer les performances de MCS, il est important de les comparer aux résultats produits par les travaux les plus récentes du domaine de la segmentation de pages Web. Il est notable de préciser que dans ce cadre, la tâche fastidieuse de codage des solutions de l'état de l'art nécessite une forte capacité d'ingénierie, et de nombreuses solutions *ad hoc* omettent des détails de mise en œuvre rendant la reproductibilité non garantie. Par conséquent, nous nous sommes souvent appuyés sur des ressources librement disponibles.

Mise en oeuvre des travaux connexes

Nous avons utilisé le *plugin Block-O-Matic* (BOM) [144] mis à disposition par les auteurs [1] pour expérimenter différentes valeurs sur un petit ensemble de pages Web représentatives ; nous avons ainsi empiriquement déterminé le meilleur paramétrage pour segmenter les 51 pages Web de notre corpus.

Nous avons également évalué l'algorithme *Box Clustering Segmentation* (BCS) [175] dont les auteurs ont aimablement partagé le code exécutable avec nous. Néanmoins, le programme fourni n'a pas été en mesure de traiter l'ensemble des 51 pages Web et seules 13 pages Web de notre ensemble de données ont pu être segmentées ; elles ont constitué la base de notre échantillon de test pour cet algorithme.

Pour concurrencer MCS, nous avons également adapté le meilleur algorithme décrit dans le chapitre précédent, GE, selon trois versions différemment optimisées. GE-D a été généralisé pour supporter une répartition équidistante de N graines initiales sur la diagonale. GE-P (notée GE-P0) à conduit à deux nouvelles versions : la première a fixé expérimentalement un seuil performant moins *ad hoc* (GE-P1) ; la deuxième utilise la méthode de pré-partitionnement en N partitions de grande taille pour identifier N graines initiales ; mais les graines une fois posées, un K-means traditionnel complet est appliqué sans tenir compte du pré-partitionnement (GE-P2). Dans tous les cas, nos choix sont limités pour GE par l'impossibilité de cette méthode à laisser libre le nombre de partitions à découvrir et faire émerger sa meilleure valeur.

Enfin, comme il est clairement montré dans [146, 145] que BOM surpasse les algorithmes *Vision-based Page Segmentation* (VIPS) et *Block Fusion* (BF), nous ne proposons pas l'intégration de ces travaux dans nos expérimentations.

1. http ://bom.ciens.ucv.ve/get-it/

Tâches et métriques d'évaluation

MCS a été évalué sur deux tâches différentes, à savoir (1) avec un nombre de partition non prédéfini à découvrir et (2) un nombre de partitions à découvrir fixé à 5. Nous utilisons l'ensemble des 51 pages Web segmentées par des experts en un minimum de partitions les plus cohérentes possibles (le résultat fut compris entre 3 et 8), et le même ensemble de données segmenté en exactement 5 partitions.

Des indices de validation interne correspondant à nos 4 fonctions objectifs permettaient de quantifier une performance de séparation et de compacité ; des indices de validation externe étaient également considérés par comparaison avec la vérité terrain proposée par les experts. Il s'agissait de 8 métriques exploitée par [5, 112] : la pureté (P), la pureté inverse (INP), l'indice de Rand (RI), l'indice de Rand ajusté (ARI), le coefficient de Jaccard (J), le coefficient de Folks et Mallows (F&M), le F-score (F) et la F-mesure cubique B (F_{b^3}). Notez que F_{b^3} est la seule métrique qui s'attaque à toutes les contraintes formelles énoncées par les auteurs, et désignera la métrique principale pour soutenir nos conclusions.

Résultats pour un nombre variable de partitions en sortie

Afin d'évaluer le processus évolutif, nous avons ajouté un certain nombre d'algorithmes « contrôles » tels que le classique K-means et une variante de l'algorithme MCS pour lequel le nombre de partitions est également donné *a priori* et seul le positionnement des graines est optimisé. Ce dernier n'est analysé que pour ses 4 meilleures configurations en termes de performance selon la mesure F_{b^3} de chacune des 4 fonctions objectifs G, T, A et C. Aussi, ces formes contrôles sont notées par la suite K-GTAC, K-TGAC, K-AGTC et K-CATG.

Indices de validation externe

Les résultats des 8 indices de validation externe sur l'ensemble de 51 pages Web segmentées manuellement en un nombre variable de partitions sont donnés dans le tableau 8.5.

Le premier résultat notable nous semble provenir de la **meilleure performance de la métrique F_{b^3} pour l'ensemble des configurations de MCS testées** comparé à tous les travaux connexes, et en particulier comparé aux meilleures solutions GE-PX issues de nos travaux précédents.

En observant les résultats, en particulier au regard de la forme contrôle K-means, nous déduisons que certaines configurations de MCS sont à privilégier. Les **meilleures configurations observées selon la métrique F_{b^3} sont de la forme AGXX**, X représentant n'importe laquelle des fonctions objectifs restantes. Cela confirme les résultats précédents [23, 171] indiquant le rôle majeur de l'alignement et de la distance visuelle dans la qualité de segmentation de pages Web. Les autres stratégies comparables en terme de performance sont celles qui combinent à la fois l'alignement et la minimisation du nombre de ruptures, c'est-à-dire ACXX et dans une moindre mesure CAXX ; résultat intéressant qui confirme nos travaux précédents concluant à l'impact négatif des ruptures logiques sur la segmentation.

Il est également pertinent de noter la **faiblesse de la stratégie optimisant en premier lieu**

	Algorithmes	P	INP	RI	ARI	J	F&M	F	F_{b^3}
MCS	GTAC(*)	0.730	0.804	0.736	0.432	0.486	0.650	0.672	0.681
	TGAC(*)	0.726	0.798	0.718	0.407	0.459	0.632	0.654	0.655
	AGTC	**0.790**	**0.913**	0.823	0.619	0.630	**0.772**	**0.769**	**0.772**
	AGCT	**0.790**	**0.913**	0.823	0.619	0.630	**0.772**	**0.769**	**0.772**
	ATGC	0.776	0.907	0.807	0.590	0.613	0.758	0.755	0.757
	ATCG	0.776	0.907	0.807	0.590	0.613	0.758	0.755	0.757
	ACGT	0.789	**0.913**	**0.825**	**0.620**	**0.631**	**0.772**	**0.769**	0.771
	ACTG	0.789	**0.913**	**0.825**	**0.620**	**0.631**	**0.772**	**0.769**	0.771
	CGTA	0.749	0.862	0.770	0.508	0.550	0.704	0.719	0.725
	CGAT	0.749	0.862	0.770	0.508	0.550	0.704	0.719	0.725
	CTGA	0.762	0.880	0.782	0.532	0.563	0.719	0.731	0.736
	CTAG	0.762	0.880	0.782	0.532	0.563	0.719	0.731	0.736
	CAGT	0.781	0.898	0.807	0.583	0.604	0.751	0.756	0.760
	CATG	0.781	0.898	0.807	0.583	0.604	0.751	0.756	0.760
Travaux connexes	BOM	0.609	0.852	0.620	0.258	0.410	0.595	0.600	0.597
	BCS(**)	0.651	0.760	0.593	0.206	0.375	0.558	0.571	0.569
	GE-P0	0.722	0.649	0.711	0.304	0.364	0.532	0.594	0.606
	GE-P1	0.670	0.798	0.650	0.282	0.395	0.584	0.603	0.623
	GE-P2	0.633	0.825	0.600	0.238	0.395	0.583	0.579	0.601
	GE-D	0.682	0.647	0.706	0.295	0.373	0.536	0.597	0.576
MCS contrôle	K-means	**0.815**	0.746	0.804	0.522	0.518	0.676	0.721	0.703
	K-GTAC	0.811	0.733	0.777	0.465	0.472	0.642	0.701	0.693
	K-TGAC	0.771	0.751	0.757	0.435	0.470	0.636	0.683	0.675
	K-AGTC	**0.850**	0.818	**0.840**	**0.625**	0.618	0.759	**0.787**	0.768
	K-CATG	0.835	0.788	0.818	0.569	0.572	0.721	0.756	0.746

TABLE 8.5. – Évaluation du partitionnement par indices de validation externe sur les 51 pages web constituant la vérité terrain et segmentée en un nombre libre de partitions ($K = [3..8]$). (*) Toutes les combinaisons de sélection commençant par cette lettre donnent les mêmes résultats. (**) Les résultats ont été calculés à l'aide de la boîte à outils de [175], mais certaines erreurs de rendu étaient présentes et seules 13 pages Web ont pu être segmentées.

le contenu sémantique textuel (TGAC*). Cela pourrait souligner la primauté des structures visuelles sur la cohérence sémantique dans la tâche imposée aux experts favorisant le *first glance* ; toutefois sans remettre en cause notre hypothèse de lien indéfectible entre sémantique et morpho-disposition.

L'examen du groupe contrôle indique une supériorité de AGTC sur K-AGTC (pour lequel le nombre de partitions est fourni en entrée de l'algorithme) avec la métrique F_{b^3}. La diversité des solutions avec différentes valeurs de K semble améliorer le positionnement correct des graines. De plus, la comparaison de cette configuration au K-means nous permet d'affirmer que le processus fournit de moins bons résultats avec un positionnement aléatoire des graines initiales.

Ces conclusions s'appuient uniquement sur l'analyse de la métrique F_{b^3} qui satisfait la plupart des contraintes imposées au partitionnement [5]. Cependant, les 7 autres métriques permettent l'analyse de sous-ensembles de ces contraintes. Même si les configurations AGXX et ACXX restent comparables et plus performantes que toutes les autres, la pureté (P) est par

exemple un cas intéressant. Dans son cas, en pénalisant les partitions non homogènes, et donc en surévaluant les solutions composées d'un grand nombre de petites partitions, cette métrique privilégie nos versions de contrôle.

Indices de validation interne

Algorithmes		DBV ↓	DBT ↓	SIA ↑	Cuts ↓	ANC
MCS	GTAC(*)	**0.76**	6.29	0.867	1.60	3.86
	TGAC(*)	2.75	**3.67**	0.873	2.46	4.52
	AGTC	1.83	5.41	**0.954**	0.50	3.56
	AGCT	1.83	5.41	**0.954**	0.50	3.56
	ATGC	2.12	5.16	**0.954**	0.70	3.58
	ATCG	2.12	5.16	**0.954**	0.70	3.58
	ACGT	1.86	5.36	**0.954**	0.46	3.52
	ACTG	1.86	5.36	**0.954**	0.46	3.52
	CGTA	1.28	6.37	0.912	**0.10**	3.40
	CGAT	1.28	6.37	0.912	**0.10**	3.40
	CTGA	1.85	5.12	0.921	**0.10**	3.48
	CTAG	1.85	5.12	0.921	**0.10**	3.48
	CAGT	1.57	6.14	0.942	**0.10**	3.40
	CATG	1.57	6.14	0.942	**0.10**	3.40
Travaux connexes	BOM	68.0	**3.21**	0.925	1.92	3.35
	BCS(**)	4.45	3.43	0.796	2.41	3.58
	GE-P0	2.92	8.18	0.759	2.45	4.67
	GE-P1	5.05	3.36	0.864	3.10	4.69
	GE-P2	15.5	4.45	0.869	3.41	4.69
	GE-D	7.67	3.97	0.694	5.67	4.69
MCS contrôle	K-means	8.11	6.54	0.823	2.80	5.50
	K-GTAC	0.95	6.75	0.816	2.26	5.50
	K-TGAC	46.0	3.94	0.865	2.74	5.50
	K-AGTC	6.23	5.91	0.932	1.28	5.50
	K-CATG	5.64	6.13	0.889	0.40	5.50

TABLE 8.6. – Évaluation du partitionnement par indices de validation interne sur les 51 pages web constituant la vérité terrain et segmentée en un nombre libre de partitions ($K = [3..8]$). (*) Toutes les combinaisons de sélection commençant par cette lettre donnent les mêmes résultats. (**) Les résultats ont été calculés à l'aide de la boîte à outils de [175], mais certaines erreurs de rendu étaient présentes et seules 13 pages Web ont pu être segmentées.

Les indices de validation interne permettent une analyse plus qualitative du partitionnement. Outre nos quatre objectifs, nous indiquons également pour chaque configuration le nombre moyen de partitions découvertes (colonne ANC). Les résultats sont donnés dans le Tableau 8.6.

Logiquement, chaque objectif favorise les configurations l'optimisant en priorité (GXXX sont les plus performantes sur DBV, TXXX, sur DBT, AXXX sur SIA at CXXX sur Cuts).

Le nombre moyen de partitions à découvrir dans nos groupes contrôle est de 5.5 et TXXX s'en rapproche le plus ; cette configuration étant par ailleurs la moins performante selon les indices de validation externes, il est probable que ces partitions soit numériquement plus optimales

mais mal formées sur d'autres paramètres. De manière générale la **tendance de MCS à produire moins de partitions** soulève un nouveau défi pour améliorer ce paramètre sans dégrader les performances.

Notons également que les configurations AXXX et CXXX, les meilleures du point de vue des indices de validation externe, sont bien plus hétérogènes sur les indices de validation interne, à l'exception de AGXX et ACXX qui présentent des comportements de partitionnement similaires.

Une analyse plus spécifique nous permet d'associer des caractéristiques aux algorithmes selon la métrique observée. Par exemple, BOM est le moins performant sur l'indice DBV, indiquant un partitionnement très peu géométrique ; tout en étant le meilleur sur DBT, faisant de cette approche un bon compromis si l'objectif est orienté prioritairement vers la cohérence sémantique. Cela n'était pas une évidence *a priori* dans l'approche qui le sous-tend, essentiellement tournée vers l'analyse du DOM. De même, la Comparaison des GE entre eux indique également que leur classement dépend des objectifs envisagés. Enfin, les résultats montrent clairement que les configurations MCS sont plus stables que les travaux connexes pour les 51 pages Web sur le critère des ruptures logiques.

Expérience avec un nombre de partitions fixé à 5

Algorithmes		P	INP	RI	ARI	J	F&M	F	F_{b^3}	Cuts
MCS	GTAC	0.813	0.733	0.783	0.487	0.493	0.657	0.702	0.697	2.08
	TGAC	0.751	0.742	0.742	0.410	0.449	0.616	0.668	0.659	2.88
	AGTC	**0.835**	**0.816**	**0.819**	**0.586**	**0.587**	**0.735**	**0.768**	**0.757**	1.12
	CATG	0.830	0.761	0.796	0.526	0.530	0.685	0.730	0.729	**0.41**
Travaux connexes	GE-P0	0.762	0.652	0.737	0.366	0.398	0.567	0.619	0.633	1.94
	GE-P1.	0.689	0.764	0.669	0.293	0.394	0.570	0.612	0.630	3.0
	GE-P2.	0.722	0.820	0.697	0.370	0.452	0.629	0.658	0.681	2.21
	GE-D	0.796	0.739	0.776	0.471	0.482	0.648	0.708	0.694	1.46
	GE-Z.	0.747	**0.829**	0.753	0.470	0.520	0.686	0.711	0.703	2.04
	GE-F.	0.715	0.795	0.707	0.373	0.450	0.624	0.662	0.667	1.88
	K-means	0.806	0.738	0.787	0.495	0.499	0.662	0.709	0.693	2.31

TABLE 8.7. – Résultats par indices de validation externe et interne sur les 51 pages Web constituant la vérité terrain et segmentées manuellement en 5 partitions. Seules les meilleures configurations par rapport à F_{b^3} de la première expérience (i.e. $K = [3..8]$) ont été prises en compte pour MCS.

La deuxième expérience a consisté à évaluer MCS sur l'ensemble des 51 pages web segmentées manuellement en un nombre fixe de 5 partitions. Le processus évolutif de MCS a été paramétré pour trouver le positionnement optimal des graines d'exactement 5 chromosomes. La taille des vecteurs représentant les noeuds des cartes auto-organisatrice de Kohonen a également été fixée à 5.

En particulier, nous avons testé les meilleures configurations de MCS pour la métrique F_{b^3} de la première expérience ; c'est à dire les solutions GTAC, TGAC, AGTC et CATG. Nous les avons comparées aux algorithmes de la série GE (D, F, Z, P0, P1 et P2) et avec la forme contrôle

K-means. Les résultats globaux pour les indices de validation externe et interne sont illustrés dans le tableau 8.7. Notez que seul le nombre de coupes est donné comme indice de validation interne.

Les valeurs de performance montrent clairement que MCS, dans sa configuration AGTC, surpasse toutes les autres stratégies de partitionnement pour 7 indices de validation externe sur 8. Seuls les algorithmes GE Z. et GE-P2 présentent un résultat légèrement meilleur pour la Pureté Inverse (INP) ; cette dernière favorisant les partitions plus équilibrées. La deuxième meilleure performance provient de la configuration CATG pour 6 des 8 indices de validation externe. Ces résultats confirment les conclusions initiales selon lesquelles les meilleures configurations favorisent l'alignement et un faible nombre de ruptures, c'est-à-dire les indices visuels et logiques.

Cependant, contrairement à la première expérience, les variantes GTAC et TGAC ne sont pas compétitives par rapport au meilleur algorithme GE, ce qui montre que l'indice sémantique textuel est la caractéristique la moins discriminante de notre expérience. L'autre résultat opposé au cas précédent concerne les variantes de GE qui n'incluent pas le pré-partitionnement ; celles-ci ont de meilleures performances que celles qui l'incluent.

En ce qui concerne le nombre de ruptures logiques, la configuration AGTC montre le deuxième meilleur résultat global, seulement dépassé par la variante naturelle pour cet indice (CATG).

Bilan conclusif

Bien que des résultats concluants aient pu être obtenus dans le cadre de cette série d'expériences, un grand nombre de directions de travail futures peuvent être proposées. Tout d'abord, d'autres expériences devraient être menées sur différents ensembles de données, en s'attaquant éventuellement à différentes langues.

Deuxièmement, comme les caractéristiques sémantiques textuelles semblent être les moins discriminantes, des modèles plus puissants peuvent être utilisés, tels que des modèles de langue basés sur des modèles auto-attentifs (transformeurs) spécifiquement adaptés, comme BERT [48] ou CamenBERT pour la langue française [99]. Des caractéristiques de densité du texte pourraient également être introduites comme le propose [77] dans l'algorithme de fusion de blocs (BF). Les avancées sur les cartes de plongements textuel [172] qui combinent les informations visuelles et textuelles dans un espace latent pourraient également constituer une direction de recherche intéressante.

Troisièmement, bien que les résultats actuels montrent des performances élevées pour les indices de validation externes, le nombre de partitions découvertes reste inférieur à celui attendu. Pour surmonter ce problème, des objectifs ou des paramètres supplémentaires pourraient être définis et introduits.

Quatrièmement, MCS peut être adapté à d'autres principes algorithmiques. Notre problèmatique de nature multi-critères pourrait inclure des versions multi-vues de K-means comme proposé dans [35]. Une manière d'interpréter nos résultats peut également conduire notre capacité à pondérer chaque modalité (visuelle, logique, textuelle) pour mieux prendre en compte leur implication dans le processus de partitionnement. Enfin, des études supplémentaires devraient

être réalisées pour sélectionner automatiquement la meilleure solution du front optimal de Pareto par une étape supplémentaire de méta-partitionnement, comme proposé dans [113], pour établir un compromis entre toutes les solutions proposées.

Outre ce bilan, l'intégration de la sémantique morho-dispositionnelle dans les problématiques de traitement automatique des langues et plus généralement d'interfaces langagières doit continuer d'être interrogée. C'est le sens du fil conducteur que j'ai essayé de suivre depuis mes premiers travaux académiques à Toulouse puis à Caen. La dernière partie conclusive de ce document s'attache à mettre en perspectives cette démarche à travers les projets et encadrements en cours de mise en place.

Perspectives et Conclusions

Contexte et environnement :

- *Laboratoire de recherche en sciences du numérique de Caen (GREYC)*
- *Équipe IMAGE depuis 2021*
- *Maître de conférence Hors Classe - 3 thèses en co-encadrement*

Après avoir survolé dans les deux parties précédentes 20 années de travaux scientifiques, et leur convergence vers mes problématiques actuelles, il reste à explorer la manière dont ils soutiendront la cohérence des futurs projets. Mes centres d'intérêt s'inscriront transversalement dans trois cadres de travail qui circonscriront mes actions scientifiques et l'organisation de cette partie conclusive :

- un cadre applicatif autour de la e-santé et du handicap sensoriel ou cognitif (non-voyance, pathologies psychiatriques ou sociales) dans des tâches spécifiques (lecture, recherche d'information) ;
- un cadre collaboratif pluridisciplinaire (informatique de l'image et du langage, linguistique, psychologie, sciences sociales, neurosciences) orienté à l'intersection du traitement automatique des langues et de l'interaction homme machine ;
- un cadre de recherche en intelligence artificielle, tourné vers la combinaison d'approches statistiques plus contrôlables, et la transposition, sans pertes informationnelles ou cognitives, d'une ou plusieurs modalités de présentation vers une ou plusieurs autres.

Ces trois cadres de travail me permettrons de développer l'avenir scientifique, technique et industriel des projets TactiNET et TagThunder et, au-delà, mes questionnements de recherche à court et moyen terme qui sont susceptibles d'en découler :

- Comment exploiter les documents dans toutes leurs dimensions ? Ou la question de la sémantique morpho-dispositionnelle appliquée à la complexité des pages Web ;
- Comment découper une page Web en zones d'intérêt de lecture ? Ou la question de l'accès global aux documents ;
- Comment représenter une zone d'intérêt de lecture par ses points saillants ? Ou la question de l'accès local aux documents ;
- Quelles influences mutuelles entre les relations rhétoriques des zones de lecture et les relations sémantiques des contenus ? Ou les questions du rapport texte/image et de l'inclusion sémantique des unités lexicales ;

- Comment transposer efficacement une structure visuelle complexe en paysage tactile et/ou sonore ? Ou la question de la définition des contours d'une " théorie gestaltiste non visuelle " ;
- Comment interagir de manière autonome à la fois localement et globalement avec les spécificités perceptives de " micromondes " tactiles et/ou sonores ? Ou les questions de l'accessibilité numérique et des interfaces langagières énactives.

Aussi, je conclurai en décrivant comment les recherches que je souhaite diriger traverseront ces perspectives par les nouveaux projets et encadrements déjà engagés.

Continuums textuels : du pictogramme au texte savant

Nous avons dédié la sous-section 1.2.1 à notre point de vue sur l'ancrage de l'évolution de l'écriture dans une petite histoire de l'autonomie ; d'abord de l'écrit sur l'oral, puis du lecteur sur le rédacteur voire de la lecture sur le texte lui même. Nous avons d'ailleurs décrit un peu plus loin, dans la section 5.2, notre volonté d'(en)acter cette évolution en l'intégrant dans nos recherches avec la volonté de préserver la capacité d'auto-détermination de tous les utilisateurs.

D'autre part nous pensons avoir démontré tout au long du chapitre 2 un lien indéfectible entre sémantique et morpho-disposition des textes ; émergence de ce processus quasi darwinien de transformation des supports et des usages pour l'accès interactif aux contenus multimodaux des documents numériques.

Un temps en repos, ces problématiques d'accessibilité et d'autonomie surgissent de nouveau à travers les préoccupations d'appels à projets récents et dans lesquels nous souhaitons impulser nos approches originales.

Autonomie, lecture interactive et sémantique morpho-dispositionnelle

Par exemple, le défi 4 de l'Appel à Manifestation d'Intérêt 2021-2026 du CNRS intitulé « vieillissement et situations de handicap » [2] nous enjoins spécifiquement à « étudier la conception, la réception et l'usage des dispositifs et expérimentations innovants en matière de compensation, suppléance, d'adaptation de l'environnement et d'accompagnement humain des personnes en vue de leur autonomie ». Ce regain d'intérêt nous semble en partie imputable à une nouvelle étape dans l'évolution des supports d'accès à un Web toujours plus multi-support, multi-application, multitâche, multi-objet et déportant une partie de la complexité visuelle et organisationnelle de l'information vers une complexité d'interactions et d'auto-adaptations des interfaces.

Nous l'avons vu, lorsqu'elles sont bien calibrées et clairement perceptibles, les propriétés typographiques et dispositionnelles permettent à un utilisateur habitué de prélever rapidement un grand nombre d'indices, et d'activer des stratégies interactives non linéaires cohérentes avec

2. https ://anr.fr/fr/detail/call/autonomie-vieillissement-et-situations-de-handicap-avh-appel-a-manifestation-dinteret-2021/

ses objectifs de lecture. En revanche, certaines populations d'usagers peuvent être mises en difficulté par une dégradation des capacités de décodage sémantique, logique ou visuel de ces contenus fortement dynamiques et structurés et des possibilités interactives associées. Cette réflexion conduit le projet d'élargissement à la fois de nos cadres de recherche, de collaboration et d'application.

projet CETELEM : CEcité TExtuelle et LEcture Multigrain

Nous regrouperons sous le terme de « cécité textuelle » cette inaptitude à interpréter facilement les documents en raison de contraintes physiologiques (malvoyance, non voyance), cognitive (dysphasie réceptive, handicap mental) ou situationnelles (petits écrans, regard occupé). Dans une perspective de conception universelle, ce cadre généralisé combine des intérêts théoriques, applicatifs et méthodologiques : les solutions apportées à une des populations d'usagers peuvent faire l'objet d'évaluations et d'adaptations dirigées vers les autres catégories d'utilisateurs concernées.

Notre projet a pour ambition de développer et exploiter les résultats obtenus et en cours d'acquisition dans les projets TactiNET et TAGTHUNDER pour les étendre aux usagers souffrant plus largement de « cécité textuelle ». En particulier, les troubles mentaux sont des problèmes majeurs de santé publique avec des enjeux sociétaux de premier plan. Deux millions de personnes sont porteuses en France d'un handicap sévère et 650 000 à 700 000 d'entre elles se trouvent en situation de handicap mental, ce qui représente environ 20 % des personnes handicapées ayant des troubles cognitifs sévères. Chaque année, entre 6 000 et 8 500 enfants naissent avec un handicap mental et les coûts générés sont estimés à 109 milliards d'euros. Ces difficultés s'accompagnent souvent d'altérations cognitives sévères déréglant le fonctionnement social et entravant l'utilisation des outils numériques et l'accès à la compréhension de la toile. Aussi, en plus de perfectionner notre solution à destination des usagers non-voyants, l'évolution envisagée concerne l'adaptation de nos algorithmes de traitement automatique des pages Web à des patients présentant des handicaps psychiatriques dégradant leur capacité de lecture silencieuse : (1) usagers avec une bonne compréhension orale mais des grandes difficultés dans la tâche de lecture ; (2) usagers dans l'incapacité d'appréhender des concepts trop abstraits (lexique inapproprié, filage de métaphore…) ou une complexité trop accrue (mise en relation des illustrations et des textes, multiplicité des échelles de lecture…).

L' architecture modulaire développée par le partenaire GREYC dans le projet TagThunder [3] sous la forme d'un enchaînement de Web Services est actuellement en phase d'expérimentation en collaboration avec l'Institut des Jeunes Aveugles et le laboratoire de recherche sur les déficiences visuelles et les technologies d'assistance de Toulouse « Cherchons pour voir » [4]. Chacun des modules intègre les différents algorithmes développés dans une perspective dirigée par la tâche de lecture et la population non-voyante visée. Nous souhaitons exploiter l'architecture existante pour adapter cette même tâche à des publics différents et faire reculer en même temps les limites posées par certaines situations de handicaps visuels et mentaux pour l'accès

3. https ://tagthunder.greyc.fr/demo/
4. https ://cherchonspourvoir.org/

aux documents ayant une complexité lexicale et structurelle inadaptée. Il s'agira d'être capable de produire à partir des textes des stimuli à un grain adaptable en terme de tâches (de communication, de vulgarisation, de recherche d'information, etc.), de complexité (littéralité, longueur, vocabulaire, etc.) et de modalité (pictogrammes, images, textes, sons, etc.).

Des collaborations centrées autour du document sont actuellement nouées pour construire ces continuums de transformation des textes, auto-contrôlables par l'usager lui même en fonction de son environnement d'interaction, de ses contraintes sensorielles et cognitives et de ses visées applicatives. Parallèlement, je participe à 3 encadrements de thèses qui démarrent pour nourrir scientifiquement cette ambition et préparer ce projet de long terme.

Approches contrastives pour les transmodalités Image/Texte/Son

Deux thèses devraient contribuer à des avancées sur ce sujet. Tournées vers le domaine du traitement automatique des langues elles abordent cette problématique de transformation de l'information depuis une modalité vers une autre selon une approche que nous qualifions de contrastive ; à l'instar de la mise en forme dont la fonction est de générer des contrastes visuels auxquels nous donnons une portée sémantique, il s'agira plus généralement de produire des contrastes (logiques, textuels, picturaux, sémantiques, etc.) appropriés à (1) un grain de lecture de la page Web, (2) une modalité cible et (3) une étape donnée de notre architecture modulaire (segmentation, extraction ou vocalisation).

Génération contextualisée de contenus multimodaux et multigrains pour l'amélioration de l'accessibilité du web

Cette thèse démarrant à la rentrée 2022, par sa nature CIFRE et la collaboration avec la société *Koena*[5], aura un double objectif : (1) la création d'un outil d'audit automatique capable de prendre en compte le contexte d'un document Web afin d'améliorer la tâche d'analyse et de générer des propositions d'améliorations cohérentes prenant en compte la sémantique morpho-dispositionnelle du document ; et (2) enrichir l'état de l'art en apprentissage d'espaces de représentation et de génération de contenus multimodaux (ici texte et image), notamment en prenant en compte la présence de contextes « positifs » et « négatifs » incarnant une sémantique morpho-dispositionnelle.

En particulier le travail se concentrera sur le concept de titre, un Objet Textuel qui peut sembler relativement peu étudié dans le domaine du TAL statistique au regard de sa prépondérance à tous les niveaux des documents et son intéressante variabilité fonctionnelle, hiérarchique et formelle [90]. Afin d'auditer le contenu d'une zone de page Web et d'apporter des propositions de corrections ou d'alternatives pertinentes, il est nécessaire de représenter le document à différentes échelles et de définir des méthodes de génération de contenus adaptées à ce contexte. Il

5. https ://koena.net/

s'agit d'une certaine manière de parvenir à un *captionning* multigrain permettant d'associer un segment textuel à n'importe quelle partie du document selon le grain de l'architecture textuelle que l'on observe (qu'il s'agisse de titrer le document, une image, un paragraphe, une zone, etc.).

Le travail s'appuiera sur les travaux développés précédemment dans le chapitre 8 [136, 8, 7, 73] et en extraction de structures visuelles ou logiques [58, 51, 49] pour modéliser les comportements pour analyser et regrouper les contenus informationnels, dans le cadre de l'évaluation des critères d'audit.

Pour cela, nous étudierons une méthode de plongement sémantique afin de représenter un document Web à différents grains et différentes modalités dans un espace sémantique commun. Ce système de représentation vectorielle unique devra permettre de proposer des métriques génériques de comparaison de contenus et d'évaluer leur pertinence en fonction du contexte. Les premières solutions s'orientent vers une adaptation du modèle UNITER [32] au français ; d'autres directions moins gourmandes en ressources et en ingénierie sont également envisagées : apprentissage de représentations multilingues [36, 173] couplées à des méthodes auto-supervisées de type auto-encodeur intégrées dans des modèles de fusion de données hétérogènes [10, 94].

La représentation fournie par ces plongements sera le point de départ de modèles génératifs de contenus. Il pourra être envisagé d'associer un titre à une zone du document Web particulièrement contrastée en utilisant des opérations arithmétiques sur les espaces latents [130, 162] et un jeu de contraintes pris en charge par la fonction de perte à optimiser. Nous pourrons également étudier l'utilisation de réseaux antagonistes génératifs [63] capables de contrôler la sémantique et la forme du contenu produit (textes ou images) dans les architectures non-supervisées. Enfin, s'il existe des données disponibles créées par des responsables d'audit, nous pourrons envisager de tirer partie de ces ressources dans un cadre d'apprentissage semi-supervisé.

Le modèle de génération de textes devra jouer sur différents paramètres stylistiques, syntaxiques ou même cognitifs (génération de texte variable en fonction d'une capacité de lecture et d'interaction spécifique). La création de différentes alternatives est d'une importance primordiale dans le cadre de la fonction d'audit et de formation de la société *Koena*. La diversité et la variabilité générées permettront à l'expert d'adapter ses propositions d'audit. Cette approche s'inspirera du travail en génération d'expressions faciales proposé par [161] où les différentes expressions sont représentées dans un espace continu qui permet de générer des expressions plus réalistes qu'à partir d'émotions discrétisées. Nous développerons ce concept en traitement automatique des langues à partir des travaux de [87].

En dernier lieu, il s'agira d'intégrer les modules développés et d'en évaluer la pertinence dans le cadre d'audit et de formation.

Modèles hiérarchiques du langage contrôlables pour la représentation et la génération

Le point de départ de ces travaux s'appuie sur les modèles encodeur-décodeur hiérarchiques qui ont pour objectif (1) de permettre de générer du texte à l'aide d'un mécanisme de masquage mais aussi (2) de produire des représentations du texte à différentes échelles. Contrairement à la

plupart des modèles de langage actuels, ils traitent l'information textuelle au niveau du caractère plutôt que du mot. L'objectif principal est de proposer une approche non supervisée capable de construire une représentation de l'information textuelle structurée à différentes échelles tout en gardant un maximum de visibilité et de contrôle sur les aspects génératifs du modèle.

La démarche statistique envisagée repose sur trois hypothèses principales. La première sous-tend notre approche hiérarchique du texte en le supposant structuré à différentes échelles graphiques. Dans un premier temps nous distinguerons les 3 grains classiques de caractère, mot et phrase. La seconde hypothèse pose que pour une échelle donnée, une unité prise hors contexte propre n'apporte qu'une quantité d'information limitée. La troisième hypothèse reconnaît que chacune des échelles envisagées porte deux types d'informations : (1) une information intrinsèque propre à la comparaison entre les unités, et (2) une information d'interaction relative à la manière dont les objets interagisses entres eux et se structurent les uns par rapport aux autres. Par exemple, à l'échelle du mot : le mot « mangez » peut être associé à des paramètres internes de longueur, d'orthographe, ou de similarité à d'autres mots comme manger. Ce même mot peut également se positionner selon une information relative à sa façon d'interagir avec d'autres mots. Ici, on peut raisonnablement dire qu'il y a une forte probabilité de trouver le mot « vous » dans son contexte.

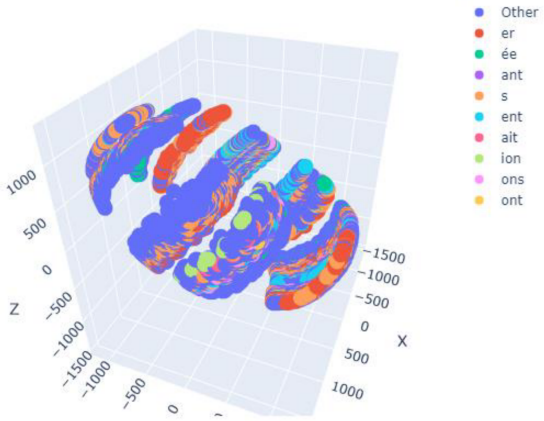


FIGURE 8.14. – Aperçu 3D de l'espace de représentation intrinsèque des mots, coloré en fonction de leur terminaison

La figure 8.14 est un exemple de vue d'une partie de l'espace de représentation que le modèle de langage hiérarchique va s'attacher à construire. On observe sa tendance naturelle à structurer l'espace de représentation en rapprochant les mots en fonction de leur terminaison. Le modèle fonctionne de manière auto-supervisée à partir de données volontairement corrompus : le texte est perturbé par un biais de masquage dépendant de l'échelle de travail (caractères, mots ou phrases) et le modèle est entraîné avec l'objectif d'apprendre à le reconstruire correctement.

Le modèle, dont les principes généraux sont décrits par la figure 8.15, utilise à chaque échelle une structure identique pour traiter l'information en deux phases. Une première phase d'encodage permet de construire deux représentations d'une unité (intrinsèque et d'interaction). Cette phase permet de passer d'une échelle à haute résolution vers une échelle à plus basse résolution,

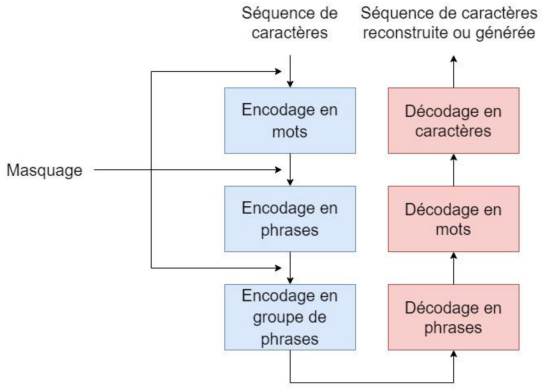


FIGURE 8.15. – Aperçu de la structure globale du modèle

par exemple de l'échelle du caractère vers l'échelle du mot. La phase suivante, de décodage, permet d'inverser le processus et de générer du texte en retrouvant des entités masquées lors de la phase d'encodage.

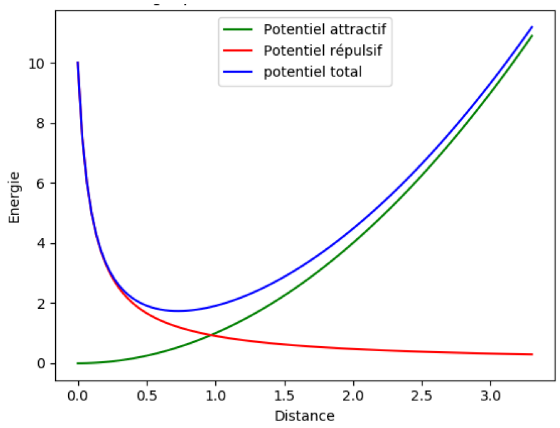


FIGURE 8.16. – Aperçu des potentiels utilisés pour structurer les espaces de représentations

La phase d'encodage utilise un mécanisme, dit « énergétique » (voir figure 8.16), pour structurer les différents espaces de représentation. Le modèle tente de minimiser une énergie construite à partir de deux potentiels : le premier provient du calcul d'une force attractive rapprochant les données similaires vis-à-vis d'une métrique donnée ; le second permet d'obtenir une force de répulsion entre tous éléments. La minimisation de ces deux potentiels permet de structurer un espace de représentation, soit (1) selon un critère de recouvrement (plus les séquences possèdent d'éléments en communs, plus les représentations sont proches) ; soit (2) selon un critère de proximité (plus les éléments au sein d'une séquence sont proches, plus les représentations de ces éléments dans l'espace sont proches).

Finalement l'objectif de ce type de modèle est de parvenir à produire des représentations reflétant de plus en plus l'aspect sémantique du texte au fur et a mesure des étapes d'encodage.

Combiné à la structuration des représentations, il devient envisageable de contrôler la génération de texte directement à l'échelle sémantique dans le but d'obtenir un résultat plus cohérent et structuré. Mais surtout, le fait de se placer à l'échelle du caractère permettra l'extension et l'étude de ce principe à l'information apportée par la sémantique morpho-dispositionnelle à tous les niveaux de notre modèle hiérarchique. Il s'agira d'intégrer de nouvelles échelles fondées par le Modèle d'Architecture Textuelle et la notion d'Objets Textuels qu'il sous-tend, dont les mots et les phrases ne sont que des représentants parmi d'autres.

Interaction multimodale et interfaces énactives

Devant l'objectif global de construire des interfaces favorisant l'auto-détermination des actions des usagers pour accéder à la sémantique morpho-dispositionnelle des documents, les travaux présentés dans la section précédente s'attachent uniquement à proposer des modèles de langage statistiques capables à la fois d'être plus multimodaux, contrôlables et hiérarchiques. Dans ce sens ils seront capables de générer du texte, des images ou des sons représentatifs d'un grain documentaire donné.

Toujours dans le cadre de l'accès à l'information dans des conditions de cécité textuelle, cette thèse en Interaction Homme Machine a pour ambition (1) d'intégrer les modèles dans l'architecture modulaire développée lors du projet Tagthunder ; et (2) de rajouter une couche interactive pour manipuler l'environnement textuel et audio en exploitant le dispositif issu du projet TactiNET.

La problématique part de la question de la mise en place de stratégies interactives haptiques et sonores permettant une transposition multimodale efficiente de la lecture visuelle vers des publics spécifiques. Nous aurons pour but l'accès non visuel à l'information an cas : (1) de situation de handicap physique ; (2) de bonne compréhension orale mais de grandes difficultés neuropsycholinguistiques pour la tâche de lecture silencieuse ; (3) d'une difficulté mentale à appréhender des concepts trop abstraits du fait d'un lexique inapproprié, ou d'un filage de métaphore par exemple ; ou (4) d'une cognition limitée par une complexité trop accrue des pages Web (par exemple pour la mise en relation des illustrations et des textes, ou l'accès à la multiplicité des échelles de lecture).

Les résultats issus des travaux précédents décrits dans les chapitres 6 et 7 nous ont convaincus que les deux modalités impliquées amélioreront substantiellement la boucle perception/action pour l'interprétation des documents Web dans une situation de cécité textuelle. la nouvelle architecture favorisera des parcours de lecture qui s'appuient sur une cohérence multigrain : le dispositif proposera une navigation multimodale orientée par des unités sémantiques combinant aussi bien des modalités haptiques, sonores, textuelles ou picturales que des structures visuelles, logiques ou thématiques.

L'architecture visée par le travail de thèse se divise en 5 modules décrits ci-après. Les parties 1, 3 et 4 ont déjà été partiellement développées pour la segmentation, l'extraction et la vocalisation dans le projet TagThunder. La thèse se focalisera donc sur les modules 2 et 5.

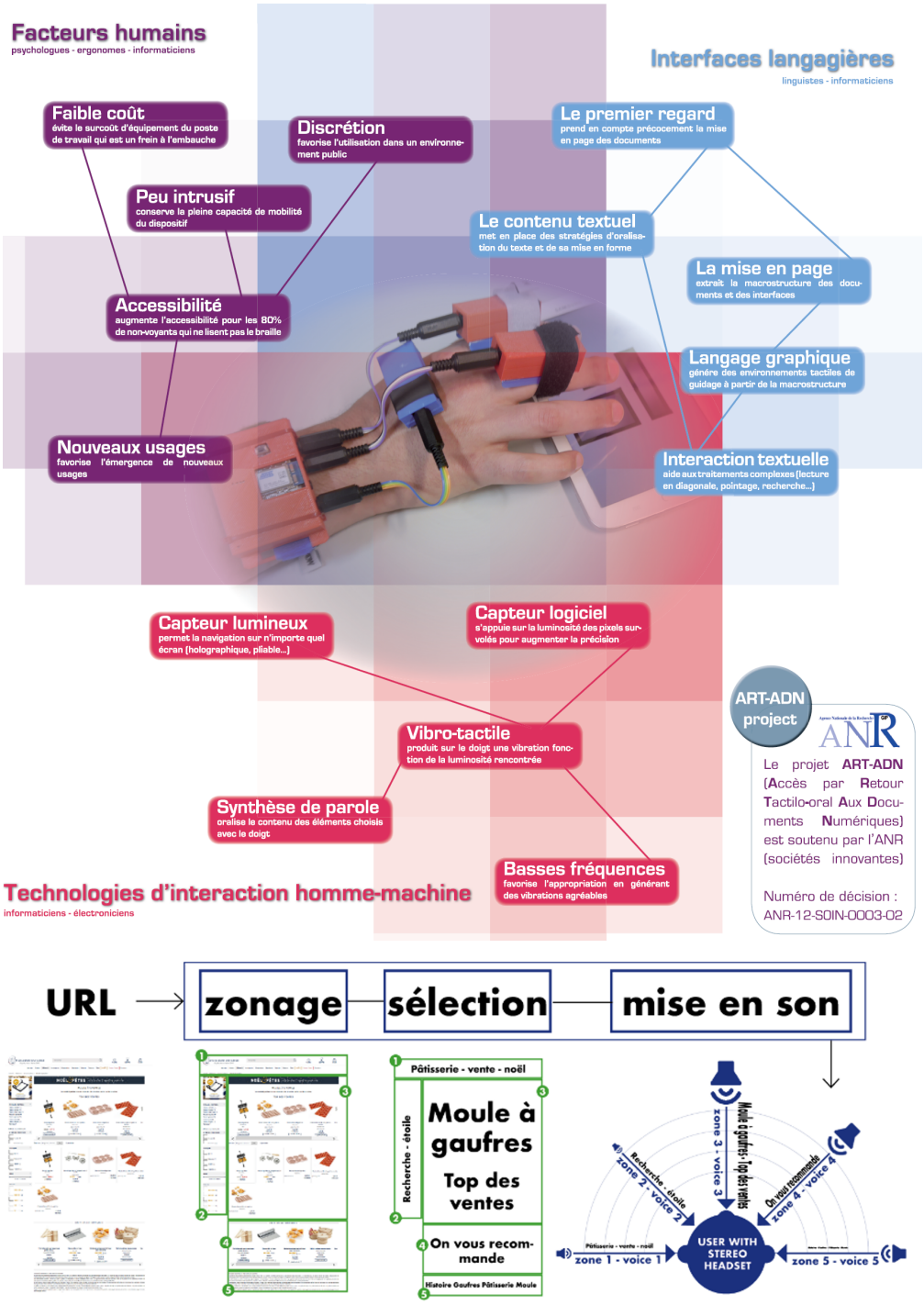


FIGURE 8.17. – Combinaison TagThunder et TactiNET

1. module de zonage : structuration des pages Web en zones cohérentes [73] ;
2. module de sélection : génération pour chaque zone des unités textuelles ou métatextuelles représentatives de la sémantique du document, mais aussi transformées pour être plus accessibles/perceptibles pour l'utilisateur envisagé ;
3. module de génération d'un paysage sonore : spatialisation sonore 3D des nuages de mots (tags cloud) formés à partir des unités textuelles préalablement extraites [83] ;
4. module de génération d'un paysage tactile : capture des contrastes de pages Web par retour tactile sur la peau (stimuli vibratoires variant en fréquence, intensité et chaleur [104]), dans le but d'appréhender la sémantique morpho-dispositionnelle des documents ;
5. module de génération d'un paysage tactilo-oral : combinaison des stimuli sonores et haptiques afin d'améliorer la boucle action/perception suivant des stratégies de lecture non visuelles. Ces nouveaux stimuli devront permettre d'interagir efficacement avec la sémantique morpho-dispositionnelle des documents.

Grâce aux travaux de cette thèse, nous souhaitons faire reculer les limites posées par certaines situations de handicaps visuels ou mentaux (cécité textuelle) pour l'accès à des documents aux complexités lexicales et structurelles fortes. Dans ce cadre, la thèse devra exploiter cette architecture pour l'augmenter d'un module supplémentaire de génération de paysages tactilo-oral : afin de transposer les modèles de lecture visuelle vers les modalités tactiles et orales, notamment dans le cadre de cécité textuelle, nous proposerons d'une part la création de tonnerres de mots dynamiques et d'autre part des modèles d'interaction multimodaux.

Création de tonnerres de mots dynamiques : a contrario des tagthunders qui sont des nuages de mots sonores statiques, la multimodalité tactilo-orale nous permet d'envisager une navigation contrôlée par l'utilisateur dans un espace sonore. Ainsi, nous proposerons que la sémantique des nuages s'adapte dynamiquement à la zone survolée par l'utilisateur. Il s'agira de mettre en cohérence la direction de navigation, le contenu sémantique survolé, celui à proximité et son grain. A notre connaissance, ce mode de transmodalisation n'a jamais été proposé dans la littérature de l'interaction homme-machine.

Modèle tactilo-oral de la lecture silencieuse : un utilisateur dispose d'une multitude de stratégies de lecture (en particulier basées sur des capacités de *skimming* et de *scanning*). Différents travaux se sont attachés d'une part à proposer des stratégies de lecture non visuelle suivant une seule modalité (soit orale, soit haptique) ou d'autre part ne considérant qu'une seule stratégie de lecture (soit *skimming*, soit *scanning*) . Dans cette thèse, nous développerons des modèles d'interaction multimodale dans une approche active où *skimming* et *scanning* peuvent s'alterner successivement de façon à optimiser la boucle action/perception.

Afin d'évaluer le prototype final nous envisagerons deux cas particuliers d'évaluation. Dans un premier temps, nous testerons le dispositif avec des utilisateurs non-voyants ou malvoyants. Pour ce faire, nous réitérerons les expériences menées dans le cadre du projet TagThunder avec l'Institut des Jeunes Aveugles de Toulouse. Nous pourrons ainsi appréhender la capacité de la nouvelle plateforme multimodale (augmentée de la couche interactive) à mieux rendre compte du contenu des documents Web et donc permettre une meilleure transposition de la lecture visuelle. Dans un deuxième temps, nous élargirons notre champ d'étude à des utilisateurs pré-

sentant des troubles cognitifs (dysphasie réceptive, handicap mental). Cette étude sera menée en collaboration avec le laboratoire PhIND UMR-S 1237 (Physiopathologie et Imagerie des Troubles Neurologiques) en lien avec le service de psychiatrie adulte du Centre Hospitalier Universitaire de Caen. Cette étude préliminaire permettra dévaluer la capacité d'auto-adaptation par l'usager des sorties d'un dispositif tel que celui de la figure 8.17, en fonction de la sévérité et du type de déficit cognitif.

Bibliographie

[1] S. Acharya, S. Saha, J. G. Moreno, et G. Dias. Multi-objective search results clustering. Dans *25th International Conference on Computational Linguistics (COLING)*, pages 99–108, 2014.

[2] F. Ahmed, Y. Borodin, A. Soviak, M. Islam, I. Ramakrishnan, et T. Hedgpeth. Accessible skimming : faster screen reading of web pages. Dans *Proceedings of the 25th annual ACM symposium on User interface software and technology*, UIST '12, pages 367–378, Cambridge, Massachusetts, USA, Oct. 2012. Association for Computing Machinery.

[3] S. Alcic et S. Conrad. Page segmentation by web content clustering. Dans *International Conference on Web Intelligence, Mining and Semantics (WIMS)*, pages 1–9, 2011.

[4] A. Aleksandrova. La portée cadrative des constructions détachées : l'exemple des portraits journalistiques. *Studii de lingvistică*, 5, 2015.

[5] E. Amigó, J. Gonzalo, J. Artiles, et F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4) :461–486, 2009.

[6] J.-J. Andrew. *Task Oriented Web Page Segmentation*. Thèse, University of Caen Normandy, 2020.

[7] J. J. Andrew, S. Ferrari, F. Maurel, G. Dias, et E. Giguet. Web Page Segmentation for Non Visual Skimming. Dans *The 33rd Pacific Asia Conference on Language, Information and Computation*, Hakodate, Japan, 2019.

[8] J. J. Andrew, S. Ferrari, F. Maurel, G. Dias, et E. Giguet. Model-driven web page segmentation for non visual access. Dans L.-M. Nguyen, X.-H. Phan, K. Hasida, et S. Tojo, editeurs, *Computational Linguistics*, pages 191–205, Singapore, 2020. Springer Singapore.

[9] M. Arabyan. Du nouveau sur le mythe des origines de la lecture silencieuse. *M. de Mattia et A. Joly (éds), De la syntaxe à la narratologie énonciative (En hommage à René Rivara), Paris, Ophrys*, page 213, 2001.

[10] N. Audebert, C. Herold, K. Slimani, et C. Vidal. Multimodal deep networks for text and image-based document classification. Dans *Conférence Nationale sur les Applications Pratiques de l'Intelligence Artificielle*, Toulouse, France, July 2019.

[11] N. Babich. *Z-Shaped Pattern For Reading Web Content*, 2017. Last access on September 2019.

[12] P. Bach-y Rita, C. C. Collins, F. A. Saunders, B. White, et L. Scadden. Vision Substitution by Tactile Image Projection. *Nature*, 221(5184) :963–964, Mar. 1969.

[13] J. Balicki. An adaptive quantum-based multiobjective evolutionary algorithm for efficient task assignment in distributed systems. Dans *13th WSEAES International Conference on Computers (ICCOMP)*, page 417–422, 2009.

[14] N. Baudouin, P. Beust, N. Chaignaud, S. Ferrari, M. Holzem, D. Jacquet, J.-P. Kotowicz, J. Labiche, S. Mauger, F. Maurel, Y. Saidali, et E. Trupin. Les interactions homme-machine : la trace en perspective. Dans *L'Homme trace, Perspectives anthropologique des traces contemporaines*, pages pp87–103. CNRS Editions, 2011.

[15] Y. Bellik. *Interfaces multimodales : concepts, modèles et architectures*. Thèse, Paris 11, 1995.

[16] A. Berliner. Atmospharenwert von drucktypen. *Ztschf. angwandte Psychol*, 17 :165–172, 1920.

[17] N. O. Bernsen. Multimodality in language and speech systems—from theory to design support tool. Dans *Multimodality in language and speech systems*, pages 93–148. Springer, 2002.

[18] L. Bing, R. Guo, W. Lam, Z.-Y. Niu, et H. Wang. Web page segmentation with structured prediction and its application in web page classification. Dans *37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, page 767–776, 2014.

[19] L. Bloomfield. *Language*. Motilal Banarsidass Publ., 1994.

[20] A. Borodin, Y. Borodin, A. Soviak, V. Ashok, S. Disfani, et I. V. Ramakrishnan. Feel-It : Personalized Audio-Tactile Web Browsing. Dans *Proceedings of the 16th Web For All 2019 Conference - Personalizing the Web, W4A 2019, San Francisco, CA, USA, May 13-15, 2019*, pages 10 :1–10 :2, 2019.

[21] A. S. Bregman. Auditory scene analysis : The perceptual organization of sound. 1994.

[22] D. Cai, X. He, Z. Li, W.-Y. Ma, et J.-R. Wen. Hierarchical clustering of www image search results using visual, textual and link information. Dans *12th Annual ACM International Conference on Multimedia (MM)*, page 952–959, 2004.

[23] D. Cai, S. Yu, J.-R. Wen, et W.-Y. Ma. Extracting content structure for web pages based on visual representation. Dans *5th Asia-Pacific Web Conference on Web Technologies and Applications*, pages 406–417, 2003.

[24] D. Cai, S. Yu, J.-R. Wen, et W.-Y. Ma. Vips : a vision-based page segmentation algorithm. Technical Report MSR-TR-2003-79, Microsoft, November 2003.

[25] N. Carbonnel, C. Valot, C. Mignot, et P. Dauchy. Etude empirique de l'usage du geste et de la parole en situation de communication homme-machine. *Travail humain (Paris)*, 1997.

[26] S. K. Card, J. D. Mackinlay, et G. G. Robertson. The design space of input devices. Dans *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 117–124, 1990.

[27] N. Catach. *La ponctuation : histoire et système*, volume 2818. Presses Universitaires de France-PUF, 1994.

[28] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, et R. Kurzweil. Universal sentence encoder. *CoRR*, abs/1803.11175, 2018.

[29] D. Chakrabarti, R. Kumar, et K. Punera. A graph-theoretic approach to webpage segmentation. Dans *17th International Conference on World Wide Web (WWW)*, pages 377–386, 2008.

[30] H. J. Chaytor. *From script to print : an introduction to medieval literature*. Cambridge University Press, 2013.

[31] Y. Chen, W.-Y. Ma, et H.-J. Zhang. Detecting web page structure for adaptive viewing on small form factor devices. Dans *12th International Conference on World Wide Web*, pages 225–233, 2003.

[32] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, et J. Liu. Uniter : Universal image-text representation learning. Dans A. Vedaldi, H. Bischof, T. Brox, et J.-M. Frahm, editeurs, *Computer Vision – ECCV 2020*, pages 104–120, Cham, 2020. Springer International Publishing.

[33] S. Choi et K. J. Kuchenbecker. Vibrotactile display : Perception, technology, and applications. *Proceedings of the IEEE*, 101(9) :2093–2104, 2012.

[34] R. W. Cholewiak et A. A. Collins. Sensory and physiological bases of touch. *The psychology of touch*, pages 23–60, 1991.

[35] G. Cleuziou, M. Exbrayat, L. Martin, et J. Sublemontier. Cofkm : A centralized method for multiple-view clustering. Dans *9th IEEE International Conference on Data Mining (ICDM)*, pages 752–757, 2009.

[36] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, et V. Stoyanov. Unsupervised cross-lingual representation learning at scale. Dans *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[37] J. Conter, S. Alet, P. Puech, et A. Bruel. A low cost, portable, optical character reader for the blind. Dans *Development of Electronic Aids for the Visually Impaired*, pages 117–125. Springer, 1986.

[38] M. Cormer, R. Mann, K. Moffatt, et R. Cohen. Towards an improved vision-based web page segmentation algorithm. Dans *14th Conference on Computer and Robot Vision (CRV)*, pages 345–352, 2017.

[39] D. Cotto. *Traitement automatique des textes en vue de la synthese vocale*. Thèse, Toulouse 3, 1992.

[40] J. Coutaz, L. Nigay, D. Salber, A. Blandford, J. May, et R. M. Young. Four easy pieces for assessing the usability of multimodal interaction : the care properties. Dans *Human—Computer Interaction*, pages 115–120. Springer, 1995.

[41] D. Dakopoulos et N. Bourbakis. Towards a 2d tactile vocabulary for navigation of blind and visually impaired. Dans *2009 IEEE International Conference on Systems, Man and Cybernetics*, pages 45–51. IEEE, 2009.

[42] C. J. Darwin, D. S. Brungart, et B. D. Simpson. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America, 114*, page 2913, 2003.

[43] D. L. Davies et D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (2) :224–227, 1979.

[44] R. De Angelis. Qu'est-ce que veut dire interpréter dans le cadre d'une herméneutique (du) numérique ? *Interfaces numériques*, 10(3), 2022.

[45] P. B. de Mareüil, P. Célérier, T. Cesses, S. Fabre, C. Jobin, P.-Y. Le Meur, D. Obadia, B. Soulage, et J. Toen. Elan text-to-speech : un système multilingue de synthèse de la parole à partir du texte. *Traitement automatique des langues*, 42(1) :223–252, 2001.

[46] F. De Saussure. *Cours de linguistique générale*, volume 1. Otto Harrassowitz Verlag, 1989.

[47] K. Deb, A. Pratap, S. Agarwal, et T. Meyarivan. A fast and elitist multiobjective genetic algorithm : Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6(2) :182–197, 2002.

[48] J. Devlin, M. Chang, K. Lee, et K. Toutanova. BERT : pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[49] A. Doucet. Logical structure extraction from digitized books. Dans *Benchmarking State-of-the-Art Systems*, page 28. World Scientific Publishing, 2017.

[50] A. Fahim, A. Salem, F. A. Torkey, et M. Ramadan. An efficient enhanced k-means clustering algorithm. *Journal of Zhejiang University-Science A*, 7(10) :1626–1633, 2006.

[51] J.-P. Fauconnier, L. Sorin, M. Kamel, M. Mojahid, et N. Aussenac-Gilles. Détection automatique de la structure organisationnelle de documents à partir de marqueurs visuels et lexicaux. Dans *Traitement Automatique des Langues Naturelles - 2014*, pages pp. 340–351, Marseille, France, July 2014.

[52] S. Ferrari, F. Maurel, P. Beust, S. Mauger, M. Holzem, N. Baudouin, E. Trupin, Y. Saidali, J. Labiche, et D. Dionisi. Pour une recherche d'information et une veille juridique interactives et socio centrées. ent énactif et veille en droit du transport. *Revue des Sciences et Technologies de l'Information-Série ISI : Ingénierie des Systèmes d'Information*, 2 :pages–17, 2012.

[53] S. Ferrari, F. Maurel, P. Beust, S. Mauger, M. Holzem, E. Trupin, Y. Saidali, J. Labiche, J.-P. Kotowicz, N. Chaignaud, D. Dionisi, D. Jacquet, et N. Baudouin. Projet d'environnement numérique énactif - Recherche d'informations et veille en droit du transport. Dans *2ème Atelier ICT (Interactions, Contextes, Traces), dans le cadre de la conférence INFORSID 2010*, page 11 pages, Marseille, France, May 2010.

[54] G. Fitzsimmons, M. J. Weal, et D. Drieghe. Skim reading : an adaptive strategy for reading on the web. Dans *ACM Web Science Conference (WebSci)*, pages 211–219, 2014.

[55] D. Frohlich. The design space of interfaces, multimedia systems. Dans *Interaction and Applications, 1st Eurographics Workshop*, pages 53–69, 1991.

[56] O. Gapenne, K. Rovira, A. Ali Ammar, et C. Lenay. Tactos : Special computer interface for the reading and writing of 2d forms in blind people. *Universal access in HCI, inclusive design in the information society*, 10 :1270–1274, 2003.

[57] J. Gardes. *Pour une grammaire de l'écrit*. Belin, 2004.

[58] E. Giguet et N. Lucas. The book structure extraction competition with the resurgence software for part and chapter detection at caen university. Dans S. Geva, J. Kamps, R. Schenkel, et A. Trotman, editeurs, *Comparative Evaluation of Focused Retrieval*, pages 128–139, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[59] N. A. Giudice, H. P. Palani, E. Brenner, et K. M. Kramer. Learning non-visual graphical information using a touch-based vibro-audio interface. Dans *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*, pages 103–110, 2012.

[60] L. H. Goldish et H. E. Taylor. The optacon : A valuable device for blind persons. *Journal of Visual Impairment and Blindness*, 68(2) :49–56, 1974.

[61] S. Goldsmith. *Designing for the disabled : the new paradigm*. Routledge, 2012.

[62] Y.-J. Gong, W.-N. Chen, Z.-H. Zhan, J. Zhang, Y. Li, Q. Zhang, et J.-J. Li. Distributed evolutionary algorithms and their models : A survey of the state-of-the-art. *Applied Soft Computing*, 34 :286–300, 2015.

[63] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, et Y. Bengio. Generative adversarial nets. Dans *27th International Conference on Neural Information Processing Systems*, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

[64] J. Goody, J. Bazin, et A. Bensa. La raison graphique : la domestication de la pensée sauvage. 1979.

[65] J. Guerreiro et D. Gonçalves. Faster text-to-speeches : Enhancing blind people's information scanning with faster concurrent speech. Dans Yesilada et Bigham [174], pages 3–11.

[66] J. a. Guerreiro. The use of concurrent speech to enhance blind people's scanning for relevant information. *SIGACCESS Accessibility and Computing*, (111) :42–46, 2015.

[67] J. Handl et J. Knowles. An evolutionary approach to multiobjective clustering. *IEEE transactions on Evolutionary Computation*, 11(1) :56–76, 2007.

[68] Z. Harris. Structures mathématiques du langage. 1971.

[69] M. L. Hawley, R. Y. Litovsky, et J. F. Culling. The benefit of binaural hearing in a cocktail party : Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, 115(2) :833–843, 2004.

[70] M. Hollins, S. Bensmaıa, et E. Roy. Vibrotaction and texture perception. *Behavioural brain research*, 135(1-2) :51–56, 2002.

[71] B. Hughes. Haptic exploration and the perception of texture orientations. 2006.

[72] Y. B. Issa, M. Mojahid, B. Oriola, et N. Vigouroux. Accessibility for the blind : An automated audio/tactile description of pictures in digital documents. Dans *2009 International Conference on Advances in Computational Tools for Engineering Applications*, pages 591–594. IEEE, 2009.

[73] S. R. Jayashree, G. Dias, J. J. Andrew, S. Saha, F. Maurel, et S. Ferrari. Multimodal web page segmentation using self-organized multi-objective clustering. *ACM Trans. Inf. Syst.*, 40(3), mar 2022.

[74] X. Jiang, Y. Hu, et L. Hang. A ranking approach to keyphrase extraction. *SIGIR'09*, 2009.

[75] Z. Jiang, H. Yin, Y. Wu, Y. Lyu, G. Min, et X. Zhang. Constructing novel block layouts for webpage analysis. *ACM Transactions on Internet Technology*, 19(3) :1–18, 2019.

[76] J. Kiesel, L. Meyer, F. Kneist, B. Stein, et M. Potthast. An Empirical Comparison of Web Page Segmentation Algorithms. Dans *43rd European Conference on IR Research (ECIR)*, 2021.

[77] C. Kohlschütter et W. Nejdl. A densitometric approach to web page segmentation. Dans *17th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1173–1182, 2008.

[78] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1) :59–69, 1982.

[79] G. Kouroupetroglou. Incorporating Typographic, Logical and Layout Knowledge of Documents into Text-to-Speech. Dans *12th European Conference on Assistive Technologies*, Vilamoura, Portugal, 2013.

[80] R. Kuber, W. Yu, et M. S. O'Modhrain. Tactile web browsing for blind users. Dans *International Workshop on Haptic and Audio Interaction Design*, pages 75–84. Springer, 2010.

[81] G. Lakoff et M. Johnson. The metaphorical structure of the human conceptual system. *Cognitive science*, 4(2) :195–208, 1980.

[82] Q. Le et T. Mikolov. Distributed representations of sentences and documents. Dans *31st International Conference on Machine Learning (ICML)*, pages 1188–1196, 2014.

[83] J.-M. Lecarpentier, E. Manishina, F. Maurel, S. Ferrari, et G. Dias. Tag Thunder : Web Page Skimming in Non Visual Environment Using Concurrent Speech. Dans *7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2016) associated to INTERSPEECH 2016*, San Francisco, United States, 2016.

[84] J. Lemarié. *La compréhension des textes visuellement structurés : Le cas des énumérations*. Thèse, Toulouse 2, 2006.

[85] C. Lenay. It's so touching" : Emotional value in distal contact. *International Journal of Design*, 4(2) :15–25, 2010.

[86] V. Lévesque et V. Hayward. Tactile graphics rendering using three laterotactile drawing primitives. Dans *2008 Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, pages 429–436. IEEE, 2008.

[87] J. Li, M. Galley, C. Brockett, J. Gao, et B. Dolan. A diversity-promoting objective function for neural conversation models. Dans *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 110–119, San Diego, California, June 2016. Association for Computational Linguistics.

[88] I. Loddo et D. Martini. The cocktail party effect. an inclusive vision of conversational interactions. *The Design Journal*, 20(sup1) :S4076–S4086, 2017.

[89] R. F. Lorch Jr., H.-T. Chen, A. A. Jawahir, et J. Lemarié. Communicating printed headings to the ear. *Ergonomics*, 59(5) :633–640, 2016. PMID : 27267653.

[90] R. F. Lorch Jr, J. Lemarié, et H.-T. Chen. Signaling topic structure via headings or preview sentences. *Psicología Educativa*, 19(2) :59–66, 2013.

[91] C. Luc. Représentation et composition des structures visuelles et rhétoriques du texte. 2000.

[92] C. Luc, M. Mojahid, et J. Virbel. Système notationnel de l'architecture textuelle par image de page. Dans *Document électronique : méthodes, démarches et techniques cognitives (Toulouse, 24-26 octobre 2001)*, pages 233–245, 2001.

[93] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. Dans *5th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, 1967.

[94] T. Magallon, F. Béchet, et B. Favre. Fusion multimodale image/texte par réseaux de neurones profonds pour la classification de documents imprimés. Dans *15e Conférence en Recherche d'Information et Applications*, Rennes, France, May 2018.

[95] J. U. Mahmud, Y. Borodin, et I. Ramakrishnan. Csurf : a context-driven non-visual web-browser. Dans *Proceedings of the 16th international conference on World Wide Web*, pages 31–40, 2007.

[96] E. Manishina, J.-M. Lecarpentier, F. Maurel, S. Ferrari, et M. Busson. Tag thunder : Towards non-visual web page skimming. Dans *18th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 281–282, 2016.

[97] N. Marquardt, M. A. Nacenta, J. E. Young, S. Carpendale, S. Greenberg, et E. Sharlin. The haptic tabletop puck : tactile feedback for interactive tabletops. Dans *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, pages 85–92, 2009.

[98] J. Martin. Cadre d'étude de la multimodalité fondé sur les types et buts de coopération entre modalités. Dans *Actes de la conférence InforMatique'94 L'interface des mondes réels et virtuels*, pages 97–106, 1994.

[99] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, et B. Sagot. CamemBERT : a tasty French language model. Dans *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7203–7219, 2020.

[100] J. Martínez, A. S. García, D. Martínez, J. P. Molina, et P. González. Texture recognition : evaluating force, vibrotactile and real feedback. Dans *IFIP Conference on Human-Computer Interaction*, pages 612–615. Springer, 2011.

[101] H. R. Maturana et F. J. Varela. *Autopoiesis and cognition : The realization of the living*, volume 42. Springer Science and Business Media, 2012.

[102] F. Maurel. *Transmodalité et multimodalité écrit/oral : modélisation, traitement automatique et évaluation de stratégies de présentation des structures "visuo-architecturale" des textes*. Thèse, Université de Toulouse, 2004.

[103] F. Maurel, G. Dias, S. Ferrari, J.-J. Andrew, et E. Giguet. Concurrent speech synthesis to improve document first glance for the blind. Dans *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 3, pages 10–17. IEEE, 2019.

[104] F. Maurel, G. Dias, W. Safi, J.-M. Routoure, et P. Beust. Layout transposition for non-visual navigation of web pages by tactile feedback on mobile devices. *Micromachines*, 11(4), 2020.

[105] F. Maurel, J. Lemarié, et N. Vigouroux. Oralisation de structures visuelles : de la lexico-syntaxe à la prosodie. Dans A. Mettouchi et G. Ferré, editeurs, *Interfaces Prosodiques (IP 2003)*, pages 137–142, Nantes, France, Mar. 2003. a.a.i. (acoustique, acquisition, interprétation).

[106] F. Maurel et W. Safi. La TactiNET : toucher le Web... Pour mieux l'entendre. Dans *27ème conférence francophone sur l'Interaction Homme-Machine.*, page w2, Toulouse, France, Oct. 2015. ACM.

[107] M. McLuhan. The gutenberg galaxy. *Toronto : University of Toronto*, 1963.

[108] Y. Meng, J. Huang, G. Wang, C. Zhang, H. Zhuang, L. M. Kaplan, et J. Han. Spherical text embedding. Dans *32nd Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 8206–8215, 2019.

[109] M. Milicka et R. Burget. Information extraction from web sources based on multi-aspect content analysis. Dans *Semantic Web Evaluation Challenges*, pages 81–92. Springer, 2015.

[110] G. A. Miller. The magical number seven, plus or minus two : Some limits on our capacity for processing information. *Psychological review*, 63(2) :81, 1956.

[111] M. Morel et A. Lacheret-Dujour. " kali", synthèse vocale à partir du texte : de la conception à la mise en oeuvre. *Traitement automatique des langues*, 42 :193–221, 2001.

[112] J. G. Moreno et G. Dias. Adapted b-cubed metrics to unbalanced datasets. Dans *38th International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, page 911–914, 2015.

[113] A. Mukhopadhyay, U. Maulik, et S. Bandyopadhyay. Multiobjective genetic clustering with ensemble among pareto front solutions : Application to mri brain image segmentation. Dans *7th International Conference on Advances in Pattern Recognition (ICPRAM)*, pages 236–239, 2009.

[114] A. Mukhopadhyay, U. Maulik, et S. Bandyopadhyay. A survey of multiobjective evolutionary clustering. *ACM Computing Survey*, 47(4), 2015.

[115] L. Nigay et J. Coutaz. A design space for multimodal systems : concurrent processing and data fusion. Dans *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 172–178, 1993.

[116] L. Nigay et J. Coutaz. Espaces conceptuels pour l'interaction multimédia et multimodale. *Technique Et Science Informatiques - TSI*, 15, 01 1996.

[117] T. C. Nogueira, D. J. Ferreira, S. T. Carvalho, et L. O. Berreta. Evaluating responsive web design's impact on blind users. *IEEE MultiMedia*, 24(2) :86–95, 2017.

[118] W. J. Ong. *Orality and literacy*. Routledge, 2013.

[119] L. Pantera, C. Hudin, et S. Panëels. Two-point haptic pattern recognition with the inverse filter method. Dans *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*, pages 545–553. Springer, 2020.

[120] L. Pantera, C. Hudin, et S. Panëels. Lotusbraille : Localised multifinger feedback on a surface for reading braille letters. Dans *2021 IEEE World Haptics Conference (WHC)*, pages 973–978. IEEE, 2021.

[121] T. N. Pappas, V. C. Tartter, A. G. Seward, B. Genzer, K. Gourgey, et I. Kretzschmar. Perceptual dimensions for a dynamic tactile display. Dans *Human Vision and Electronic Imaging XIV*, volume 7240, page 72400K. International Society for Optics and Photonics, 2009.

[122] B. Parmanto, R. Ferrydiansyah, A. Saptono, L. Song, I. W. Sugiantara, et S. Hackett. AcceSS : accessibility through simplification and summarization. Dans *Proceedings of the International Cross-Disciplinary Workshop on Web Accessibility, Chiba, Japan, May 10-14, 2005*, pages 18–25, 2005.

[123] E. Pascual. Représentation de l'architecture textuelle et génératio nde textes. 1991.

[124] K. Pernice. *F-Shaped Pattern of Reading on the Web : Misunderstood, But Still Relevant (Even on Mobile)*, 2017. Last access on September 2019.

[125] K. Pernice. Text scanning patterns : Eyetracking evidence. *Nielsen Norman Group*, 25(8), 2019.

[126] K. Pernice, K. Whitenton, et J. Nielsen. *How People Read on the Web : The Eyetracking Evidence*. Nielsen Norman Group, 2014.

[127] G. Petit, A. Dufresne, V. Levesque, V. Hayward, et N. Trudeau. Refreshable tactile graphics applied to schoolbook illustrations for students with visual impairment. Dans *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, pages 89–96, 2008.

[128] G. Petit, A. Dufresne, et J.-M. Robert. Introducing tactoweb : A tool to spatially explore web pages for users with visual impairment. Dans *International Conference on Universal Access in Human-Computer Interaction*, pages 276–284. Springer, 2011.

[129] R. Power, D. Scott, et N. Bouayad-Agha. Document structure. *Computational Linguistics*, 29(2) :211–260, 2003.

[130] A. Radford, L. Metz, et S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. Dans Y. Bengio et Y. LeCun, editeurs,

4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, 2016.

[131] J. Robertson et J. H. Nash. Mob-suite : software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial genomics*, 4(8), 2018.

[132] M. Rossi. *L'intonation : le système du français : description et modélisation*. Editions Ophrys, 1999.

[133] M. Rotard, S. Knödler, et T. Ertl. A tactile web browser for the visually disabled. Dans *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*, pages 15–22, 2005.

[134] P. J. Rousseeuw. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20 :53–65, 1987.

[135] P. Saenger. Silent reading : Its impact on late medieval script and society. *Viator*, 13 :367–414, 1982.

[136] W. Safi, F. Maurel, J. Routoure, P. Beust, et G. Dias. A hybrid segmentation of web pages for vibro-tactile access on touch-screen devices. Dans *VL@COLING*, 2014.

[137] W. Safi, F. Maurel, J. Routoure, P. Beust, et G. Dias. Web-adapted supervised segmentation to improve a new tactile vision sensory substitution (TVSS) technology. Dans *Proceedings of the 6th International Conference on Ambient Systems, Networks and Technologies (ANT 2015), the 5th International Conference on Sustainable Energy Information Technology (SEIT-2015), London, UK, June 2-5, 2015*, pages 35–42, 2015.

[138] W. Safi, F. Maurel, J.-M. Routoure, P. Beust, et G. Dias. An Empirical Study for Examining the Performance of Visually Impaired People in Recognizing Shapes through a Vibro-tactile Feedback. Dans *17th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2015)*, pages 349–350, Lisbon, Portugal, Oct. 2015.

[139] W. Safi, F. Maurel, J.-M. Routoure, P. Beust, M. Molina, C. Sann, et J. Guilbert. Which ranges of intensities are more perceptible for non-visual vibro-tactile navigation on touch-screen devices. Dans *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, pages 1–2, 2017.

[140] S. Saha et S. Bandyopadhyay. A symmetry based multiobjective clustering technique for automatic evolution of clusters. *Pattern recognition*, 43(3) :738–751, 2010.

[141] A. Sahami, P. Holleis, A. Schmidt, et J. Häkkilä. Rich tactile output on mobile devices. Dans *European Conference on Ambient Intelligence*, pages 210–221. Springer, 2008.

[142] N. Saini, S. Saha, et P. Bhattacharyya. Automatic scientific document clustering using self-organized multi-objective differential evolution. *Cognitive Computation*, pages 1–23, 12 2018.

[143] P. Salza et F. Ferrucci. Speed up genetic algorithms in the cloud using software containers. *Future Generation Computer Systems*, 92 :276–289, 2019.

[144] A. Sanoja et S. Gançarski. Block-o-Matic : A web page segmentation framework. Dans *International Conference on Multimedia Computing and Systems (ICMCS)*, pages 595–600, 2014.

[145] A. Sanoja et S. Gançarski. Web page segmentation evaluation. Dans *30th Annual ACM Symposium on Applied Computing (SAC)*, 2015.

[146] A. Sanoja Vargas. *Web page segmentation, evaluation and applications*. Thèse, Pierre and Marie Curie University, Paris, France, 2015.

[147] E. Sempère. Sylvie catellin, sérendipité. du conte au concept. paris, seuil, coll.«science ouverte», 2014, 265 p. *Féeries. Études sur le conte merveilleux, XVIIe-XIXe siècle*, (12) :172–177, 2015.

[148] V. Servais. La relation homme-animal. la relation à l'animal peut-elle devenir significative, donc thérapeutique, dans le traitement des maladies. *Enfances et Psy*, 2(35), 2007.

[149] X. Sevillano, J. C. Socoró, et F. Alías. Parallel hierarchical architectures for efficient consensus clustering on big multimedia cluster ensembles. *Information Sciences*, 511 :212 – 228, 2020.

[150] D. K. Simonton. *Serendipity and Creativity in the Arts and Sciences : A Combinatorial Analysis*, pages 293–320. Springer International Publishing, Cham, 2022.

[151] K. Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1) :11–21, 1972.

[152] J.-L. Swiners, J.-M. Briet, et E. De Bono. *L'intelligence créative au-delà du brainstorming-2e éd.* Maxima, 2017.

[153] X. Tan, T. Qin, F. Soong, et T.-Y. Liu. A survey on neural speech synthesis. *arXiv preprint arXiv :2106.15561*, 2021.

[154] P. Thagard. *How scientists explain disease*. Princeton University Press, 2018.

[155] M. Tixier, C. Lenay, G. Le Bihan, O. Gapenne, et D. Aubert. Designing interactive content with blind users for a perceptual supplementation system. Dans *Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction*, pages 229–236, 2013.

[156] B. Treutwein. Adaptive psychophysical procedures. *Vision research*, 35(17) :2503–2522, 1995.

[157] A. Tricot. Utility, usability and acceptability : an ergonomic approach to the evaluation of external representations for learning. 01 2007.

[158] M. Turgeon, A. S. Bregman, et B. Roberts. Rhythmic masking release : effects of asynchrony, temporal overlap, harmonic relations, and source separation on cross-spectral grouping. *Journal of Experimental Psychology : Human Perception and Performance, 31(5)*, page 939, 2005.

[159] R. P. Velloso et C. F. Dorneles. Extracting records from the web using a signal processing approach. Dans *2017 ACM on Conference on Information and Knowledge Management (CIKM)*, page 197–206, 2017.

[160] R. P. Velloso et C. F. Dorneles. Web page structured content detection using supervised machine learning. Dans M. Bakaev, F. Frasincar, et I.-Y. Ko, editeurs, *International Conference on Web Engineering (ICWE)*, pages 3–18, 2019.

[161] V. Vielzeuf, C. Kervadec, S. Pateux, et F. Jurie. The many moods of emotion, 2018.

[162] V. Vielzeuf, A. Lechervy, S. Pateux, et F. Jurie. Multi-level sensor fusion with deep learning. *IEEE Sensors Letters*, PP :1–1, 10 2018.

[163] J. Virbel. Structured documents. chapter The Contribution of Linguistic Knowledge to the Interpretation of Text Structures, pages 161–180. Cambridge University Press, New York, NY, USA, 1989.

[164] D. Weng, J. Hong, et D. A. Bell. Extracting data records from query result pages based on visual features. *Advances in Databases*, pages 140–153, 2011.

[165] L. Wiener, T. Ekholm, et P. Haller. Modular responsive web design : An experience report. Dans *Companion to the First International Conference on the Art, Science and Engineering of Programming*. Association for Computing Machinery, 2017.

[166] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, et C. G. Nevill-Manning. KEA : Practical Automatic Keyphrase Extraction. *NRC/ERB-1057*, 1999.

[167] J. Xin et H. Jiawei. *Quality Threshold Clustering*, pages 1–2. Springer US, Boston, MA, 2016.

[168] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, et M. Zhou. Layoutlm : Pre-training of text and layout for document image understanding. Dans *26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, page 1192–1200, 2020.

[169] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, et L. Zhou. Layoutlmv2 : Multi-modal pre-training for visually-rich document understanding, 2020.

[170] P. B. y Rita. Brain plasticity as a basis of sensory substitution. *Journal of Neurologic Rehabilitation*, 1(2) :67–71, 1987.

[171] X. Yang et Y. Shi. Web page segmentation based on gestalt theory. Dans *IEEE International Conference on Multimedia and Expo (ICME)*, pages 2253–2256, 2007.

[172] X. Yang, E. Yumer, P. Asente, M. Kraley, D. Kifer, et C. Lee Giles. Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. Dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5315–5324, 2017.

[173] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, et R. Kurzweil. Multilingual universal sentence encoder for semantic retrieval, 2019.

[174] Y. Yesilada et J. P. Bigham, editeurs. *Proceedings of the 17th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS 2015, Lisbon, Portugal, October 26-28, 2015*. ACM, 2015.

[175] J. Zeleny, R. Burget, et J. Zendulka. Box clustering segmentation : A new method for vision-based web page preprocessing. *Information Processing and Management*, 53(3) :735–750, 2017.

[176] G. Zimmermann, G. C. Vanderheiden, et C. Strobbe. Towards deep adaptivity – a framework for the development of fully context-sensitive user interfaces. Dans C. Stephanidis et M. Antona, editeurs, *Universal Access in Human-Computer Interaction. Design and*

Development Methods for Universal Access, pages 299–310, Cham, 2014. Springer International Publishing.