
Documents numériques et transmodalité

Transposition automatique à l'oral des structures visuelles des textes

**Fabrice Maurel — Mustapha Mojahid — Nadine Vigouroux
Jacques Virbel**

*Université Toulouse III
Institut de Recherche en Informatique de Toulouse
118, Route de Narbonne
F-31062 Toulouse cedex 09
{fmaurel,mojahid,vigouroux,virbel}@irit.fr*

RÉSUMÉ. Longtemps considérée comme ornementale, la structure visuelle des textes est aujourd'hui abordée comme un objet d'étude à part entière dans diverses disciplines. Nous nous intéressons à son utilité et, le cas échéant, son utilisabilité, dans le cadre de la transposition à l'oral par des systèmes de synthèse de la parole à partir de textes (TTS). Nous proposons un système d'oralisation qui conduit à des représentations du texte interprétables par un TTS. Nous focalisons sur un modèle d'oralisation par reformulation des textes écrits pour être lus silencieusement (MORTELS) dont l'objectif est de rendre « articulables » certaines parties significatives des textes souvent « oubliées » par les systèmes actuels.

ABSTRACT. Considered for a long time as ornamental, the texts visual structure becomes nowadays a full object of investigation in various disciplines. We are interested in the utility of these aspects of documents and, if the need arises, their usability, within the framework of their oral transposition by Text-To-Speech systems (TTS). We propose an Oralisation system which leads to text representations directly interpretable by TTS. We focus on a reformulation-based oralisation model for texts written to be silently read (MORTELS) in order to render some signifying parts of the texts often “forgotten” by synthesis systems.

*MOTS-CLÉS : transmodalité, synthèse de la parole à partir de textes, structures visuelles des textes.
KEYWORDS: transmodality, Text-To-Speech systems, texts visual structures.*

1. Introduction

Il est parfois fait l'hypothèse que la lecture silencieuse a eu un développement récent. Certains ont vu des raisons technologiques au déplacement d'une lecture orale vers une lecture silencieuse. L'idée que l'introduction de l'imprimerie conduit à un changement radical d'une culture orale à une culture visuelle, d'une lecture orale en groupe à une lecture silencieuse individuelle, a été suggérée par Chaytor (1945), popularisée par McLuhan (1962) et développée par Ong (1982). Pourtant la littérature indique (Gullick, 1986) que la notion de société médiévale purement orale est une simplification puisque des agencements de pages non linéaires complexes étaient courants. En fait le véritable élan de la lecture silencieuse viendrait plutôt selon Saenger (1982) des besoins fonctionnels de la croissance de la scolastique au XII^e et XIII^e siècle : 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 (on pourra voir Manguel, 1998 pour une histoire de la lecture). 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 que nous nous lançons 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 émergé grâce au fait que la lecture orale n'était justement plus de rigueur !*

Lorsque l'on considère le processus d'écriture, le texte, bien que segmenté (en mots, lignes, phrases, images...), peut voir sa linéarité cassée par l'utilisation des deux dimensions du support (en particulier pour exprimer la structure hiérarchique en sections, chapitres...). Dans ce cas, du sens peut être ajouté à la chaîne segmentée à travers la forme « analogique » de la disposition globale. De telles caractéristiques potentielles du texte (non linéaire et analogique) fournissent un défi aux modèles de la linguistique car il n'y a aucune comparaison possible entre la linéarité temporelle pure à laquelle répond l'oral, et la linéarité à la fois temporelle et spatiale (donc sur laquelle on peut revenir), qui caractérise l'écrit. Remarquons que, à l'instar des images picturales, on peut aller plus loin en arguant de l'existence de trois dimensions dans une page typographique : « Le jeu relatif des corps, des graisses, des formes et des familles permet d'accentuer les différences de gris et de structures dans une troisième dimension de l'espace visuel de la page imprimée » (Duplan, 1986). Finalement, le signifiant auditif est « une linéarité qui n'offre aucun choix, puisqu'elle dépend d'une nécessité biomécanique » (Harris, 1993), alors qu'à l'écrit la linéarité est choisie : au niveau de la direction (sens d'écriture de droite à gauche dans les langues sémitiques) et des ruptures de la continuité (le boustrophédon commande un passage à la ligne « en zigzags » alternant à gauche et à droite). Les choix se traduisent également à travers des configurations spécifiques comme par exemple les énumérations dont Luc (2000) dit qu'elles « participent à la fois de structures entièrement visuelles qui n'ont pas de source dans l'oral et de structures

entièrement discursives ». Gardes-Tamine (2004) résume ces différentes remarques en précisant que « l'écrit implique un alignement et une [ou plusieurs !] direction[s] mais pas la linéarité ». Nous exprimerons où se situe le défi que nous relevons en affirmant « [qu']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 » (Arabyan, 2001). Même si, d'un point de vue naïf, la ponctuation, préoccupation centrale dans la citation précédente, pouvait sembler être la part typographique des textes la plus aisée à mettre en correspondance avec des phénomènes oraux, nous précisons que cette absence d'imitation naturelle peut également s'appliquer en général aux autres aspects typographiques et dispositionnels du texte. Notre approche globalement pluridisciplinaire et les remarques précédentes expliquent le découpage de cet article en trois parties.

La première partie concerne la mise en évidence d'une dimension propre aux textes écrits pour être lus silencieusement qui, bien que peu traitée par des logiciels de synthèse de la parole à partir de texte, nous semble incontournable pour une oralisation « correcte ». Nous décrirons la prise en compte de cet aspect du texte à travers le point de vue des spécialistes de la synthèse de la parole à partir de textes et des linguistes.

Pour répondre aux problèmes soulevés par la première partie, nous proposons dans une seconde partie une architecture informatique basée sur une série de transformations des différents objets textuels qui composent le texte afin d'en fournir une représentation directement oralisable par les logiciels de synthèse. Nous focaliserons ensuite notre propos sur un des modules de l'architecture à travers la conception d'un modèle qui veut décrire différentes stratégies d'oralisation par reformulation de textes écrits pour être lus silencieusement (MORTELS). Ce modèle prétend mettre en relation la liste exhaustive des stratégies de transposition orale prenant en compte les informations de nature sémantique et pragmatique véhiculées par la morpho-disposition des textes.

Avant de conclure et de mettre en perspective notre travail, l'opérationnalisation de ce modèle est partiellement démontrée dans une troisième partie à travers l'exemple de l'oralisation des structures énumératives.

2. Valeur et prise en compte des structures visuelles des textes

Les systèmes de synthèse de parole à partir de textes (ou Text-To-Speech – TTS) ont pour vocation la transformation d'un texte tout-venant vers une chaîne phonétique synthétisable en une suite de sons interprétables comme de la parole humaine. Un usage fréquent de ce type de technologie est observé lorsque une grande quantité de textes dont les contenus ne sont pas prévisibles doit être rendue accessible dans une modalité non visuelle (téléphone ou déficience visuelle pour la

navigation sur le web par exemple). Pourtant, ces outils tardent à intégrer réellement la composante morpho-dispositionnelle du texte : 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, une marque ne peut être considérée isolément. Elle est un élément à part entière du texte, pouvant être liée à d'autres éléments par la constitution intentionnelle et signifiante d'une configuration particulière. L'objet des deux sous-sections suivantes est d'extraire de la littérature la place donnée à cette dimension du texte du point de vue des spécialistes de la synthèse automatique de la parole à partir de textes et des linguistes. Ce travail pourra intéresser particulièrement les chercheurs en traitement automatique des langues ou des documents, ainsi que l'industrie à travers nos discussions et recommandations autour des normes de représentation des documents écrits pour être dits.

2.1. Synthèse de la parole et morpho-disposition

Un système de synthèse de parole à partir de texte enchaîne de nombreuses procédures de calcul (Dutoit, 1997) 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, divers prétraitements spécifiques sont le plus souvent appliqués à travers un module de prétraitement du texte. 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 termes 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...). L'objectif était de modifier la chaîne orthographique afin d'obliger le synthétiseur vocal utilisé à générer la bonne

prononciation. Un exemple plus récent de prétraitement, dépendant de la langue, est donné par Torzec (2001) pour le système de synthèse de parole multilingue développé par France Telecom R&D (CVOX) qui reformulera « Au V^e siècle av. J.C., 1 kg de plastique coûtait 30576\$. » en « Au CINQUIEME siècle AVANT JESUS CHRIST, UN KILOGRAMME de plastique coûtait TRENTE MILLE CINQ CENT SOIXANTE SEIZE DOLLARS. ». Pour obtenir un tel résultat, le système s'appuie sur une analyse de surface du texte dirigée par les formes orthographiques et typographiques et des dictionnaires ; ceux-ci, à partir d'expressions régulières, permettent une segmentation du texte en entrées dites lexicales et extra-lexicales. Ces dernières sont ensuite converties dans une forme littérale appropriée à leur vocalisation (Cherbonnel, 1996). En ce qui concerne les prétraitements qui sont classifiés par les concepteurs comme indépendants de la langue, nous retrouvons en particulier la tentative d'élimination ou de conversion d'informations dites métalinguistiques (mise en forme, mise en page, etc.) qui consiste à ne transcrire que ce qui semble « intéressant » ; cela dit aucune autre précision ni référence n'est donnée sur les critères de repérage et les solutions envisagées pour la vocalisation de ces phénomènes.

D'autres exemples de prétraitement peuvent être fournis pour des systèmes sensiblement différents. Kali (Morel, 2001), logiciel de synthèse vocale à partir de texte prétraitera « l'ADN des hommes célèbres intéresse le président. » en « l'a.d.n des hommes célèbres intéresse le président », préférant ne pas développer le sigle mais plutôt désambiguïser la nature du mot « ADN » (une syllabe /adn/ versus trois syllabes /a/de/en/) en le substituant par le sigle ponctué « a.d.n ». Le système industriel multilingue commercialisé en 2001 par Elan Speech gère, lors de la phase de prétraitement, un jeu de caractères (espace, tabulation, tiret, arobase, point, virgule, deux-points, slash et apostrophe) pour permettre 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). La synthèse de parole à partir de textes a ainsi connu une évolution remarquable au cours des 30 dernières années. Son intelligibilité et son naturel ont été améliorés au point que ces deux critères de jugement deviennent de plus en plus obsolètes.

Certaines démarches scientifiques récentes rejoignent notre problématique car, en alliant le souci d'efficacité pratique et la recherche de verrous scientifiques pertinents, elles abordent plus facilement la problématique qui nous préoccupe ici. Par exemple certaines synthèses issues de ce type de recherche disposent d'un module de prétraitement de mël (Sproat, 1998), (Torzec, 2001), qui tient compte de certains phénomènes typographiques, spécifiques à ce mode de communication, se situant à un niveau plus complexe que celui de la lexico-syntaxe (par exemple : les *smileys* ou « émoticones » ; la gestion de la succession des réponses et des transferts). D'ailleurs cette problématique émergente a été relevée, dans le cadre de la présentation du système de synthèse de la parole multilingue *Elan Speech*, comme

une évolution intéressante qui nécessiterait de « déborder du cadre de la phrase, objet de la syntaxe (d'aller peut-être jusqu'au paragraphe, voire à une hiérarchisation du document, pour ne pas lire de façon linéaire, par exemple, les titres, les chapitres ou les notes de bas de page), et d'étudier dans quelle mesure il existe, dans les formes, des indices permettant de prendre en compte le sens » (Boula de Mareuil, 2001 ; p. 248).

Malgré cela, lorsque nous étudions les systèmes récents (Alessandro, 2001), nous remarquons que la mise en forme du texte reste encore peu prise en compte. De manière générale, même lorsque la linguistique devient textuelle elle s'attaque le plus souvent à la description du texte à partir d'une définition « amodale » et « désincarnée » de celui-ci. Ainsi, les différents niveaux classiques de la linguistique (lexical, syntaxique, sémantique et pragmatique) ne prennent que peu en compte les aspects spécifiques au support du texte : les modèles s'abstraient considérablement de la modalité de présentation de l'information pour expliquer les structures pragmatiques et rhétoriques du discours/texte. Toutefois, cette remarque s'applique moins au discours oral depuis qu'ont émergé la notion de structure prosodique et l'étude des relations qu'elles entretiennent avec les autres niveaux linguistiques (années 1970-1980). En revanche, peu de tentatives ont été faites pour essayer de considérer, dans le champ circonscrit par les recherches en linguistique textuelle, l'analyse de procédés spécifiques à l'écrit, pourtant tout à fait comparables aux phénomènes intonatifs : ceux relatifs à la morpho-disposition des textes. La section suivante s'attachera à mettre en valeur les exceptions notables que nous avons pu relever dans la littérature.

2.2. Linguistique et morpho-disposition

Les différents choix du rédacteur dans la gestion de la morpho-disposition peuvent influencer sur l'efficacité communicationnelle du texte mais également sur son sens. Nous affirmons en effet qu'elle participe à l'organisation sémantique des énoncés, puisque l'assemblage des marques en configurations particulières peut ajouter du sens à la chaîne alphabétique ciblée en l'encadrant, en s'y insérant ou en se superposant à elle. Un argument en faveur de cette hypothèse vient de la possibilité de montrer que la non prise en compte de ces marques est susceptible de produire des contresens :

- (1) Ce sont les linguistes qui aiment les problèmes qui sont intelligents.**
- (2) Ce sont les linguistes
qui aiment les problèmes
qui sont intelligents.**
- (3) Ce sont les linguistes, qui aiment les problèmes, qui sont intelligents.**
- (3bis) Ce sont les linguistes (qui aiment les problèmes) qui sont intelligents.**

L'interprétation de (2) est modifiée par rapport à celles de (1) alors que ces deux objets ne diffèrent que par leur spatialisation sur le support (la première phrase a deux interprétations possibles mais c'est bien une nouvelle interprétation qui émerge de la seconde phrase). La sémantique véhiculée par (2) aurait pu également être obtenue en modifiant quelque peu le texte comme le montrent (3) et (3bis). Mais c'est alors à un niveau pragmatique que la différence se fait : la structure informationnelle est modifiée en termes de focus, rhème, thème et/ou topique. Bien que la linguistique ne s'est que rarement attachée à tenir compte des implications liées au choix du support physique sur lequel la chaîne linguistique s'inscrit, nous focaliserons maintenant sur une des seules propositions, à notre connaissance, qui revêt un pouvoir explicatif plus global de la sémantico-pragmatique de la morpho-disposition des textes, à travers la notion d'architecture textuelle.

2.2.1. *Le 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 – Virbel, 1989) 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 (Harris, 1971). 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 typographiques et dispositionnelles. L'ensemble des phrases décrivant les propriétés morpho-dispositionnelles et lexico-syntaxiques forment le *métalangage architectural* (Pascual, 1991). Cette approche a pour conséquence l'existence d'un continuum entre des formes de MFM entièrement discursives et des formes visuelles : on passe d'une extrémité à l'autre de ce continuum en effaçant (au profit de marques visuelles) ou en reconstruisant (par interprétation) le métalangage architectural. L'intérêt d'utiliser cette approche dans l'étude de l'oralisation des structures textuelles est qu'elle fournit un cadre général d'interprétation et de reformulation des phénomènes typo-dispositionnels qui ne sont pas directement

transposables à l'oral. Nous allons chercher à reconstruire une partie du métalangage architectural qui est entièrement ou partiellement effacée de façon à expliciter de manière discursive les propriétés structurelles des textes.

2.2.2. Contexte de l'étude

Nous avons vu que l'architecture des systèmes de synthèse de la parole à partir de textes ne pouvaient prendre en compte une analyse sémentico-pragmatique des structures visuelles des textes à un niveau suffisant pour rendre compte convenablement du contenu informationnel qu'elles véhiculent. Selon nous, ce fait peut s'expliquer plus précisément par deux constatations (Maurel, 2004a).

D'une part l'unité d'analyse des TTS est la phrase ponctuée. Or, l'inscription morphologique des textes nécessite souvent d'être appréhendée en deça ou au-delà de cette unité pour en dégager toute la sémantique. Notre objectif est donc d'intégrer cette technologie dans une architecture informatique plus globale qui permette de fournir en bout de chaîne de traitement une représentation du texte directement oralisable par les systèmes actuels. Cela implique de savoir *mettre en relation l'inscription morphologique des textes avec une version discursive* sous la forme de suite de phrases ponctuées.

D'autre part, un langage de normalisation, le *Speech Synthesis Markup Language* (SSML) est proposé par le World Wide Web Consortium (W3C) afin de permettre de piloter la prosodie produite par les TTS à travers un ensemble de marques intégrables directement dans le texte fourni en entrée du système. Si cette possibilité paraît intéressante, elle est difficilement exploitable dans notre cadre d'étude sans la *mise en relation de l'inscription morphologique des textes avec une configuration prosodique*, ou plus généralement « typophonique », adaptée.

Le modèle d'architecture textuelle précédemment évoqué nous permet d'envisager quelques réponses à la première problématique puisqu'il nous indique une unité de traitement plus appropriée, l'objet textuel (OT), et une méthode pour recomposer le texte par des versions discursives de ces OT :

- *objet textuel* : un objet textuel (OT) est un 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 ;

- *métaphrase* : une métaphrase est une 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.

De plus, l'hypothèse d'une équivalence fonctionnelle entre typographie, disposition et lexico-syntaxe inspire notre approche pour aborder le second point. En effet, nous émettons l'hypothèse d'une équivalence fonctionnelle entre marques lexico-syntaxiques et prosodiques : certains segments de texte à portée

métalinguistique pourront être effacés au profit de configurations intonatives équivalentes. La partie suivante s'attache à proposer des solutions concrètes à ces deux solutions. La première sous-section décrit une architecture informatique intégrant un TTS plus adaptée à la transposition à l'oral des structures visuelles des textes ; la seconde se concentre sur un des modules de cette architecture et les ressources qu'il nécessite : le module de reformulation et les stratégies d'oralisation associées.

3. Architecture et modèle

3.1. *Intégration des TTS dans une architecture adaptée*

L'originalité du système proposé (Maurel, 2004b) réside dans l'unité d'analyse considérée pour le traitement des documents écrits. En effet, elle ne correspond plus à la traditionnelle phrase ponctuée et n'a pas non plus l'ambition d'intégrer le texte globalement avec son cotexte et ses contextes. En fait, le processus d'oralisation décrit par cette architecture permet de faire « traverser » à un objet textuel plusieurs phases de transformation qui s'appuient sur un ensemble de connaissances associées à chacune d'elles. Trois phases permettent, à partir d'un objet textuel « écologique », de fournir une représentation, associée à une stratégie d'oralisation particulière, interprétable par un TTS.

Les différentes étapes de ce processus sont décrites ci-après :

- phase de *segmentation* : cette étape s'appuie sur des indications d'ordre syntaxique, pragmatique, stylistique et structurel, regroupées sous le terme de base de règles éditoriales. Cette première étape permet de contrôler en partie les phénomènes de variabilité du texte « tout-venant » en reconnaissant et normalisant l'OT en entrée du processus. La segmentation de l'OT produira sa décomposition en un ensemble d'OT de niveau juste inférieur ;

- phase d'*extraction* : la deuxième étape consiste à extraire des OT obtenus les différents paramètres formels, qui sont des connaissances structurelles et linguistiques, dont la variation semble a priori pertinente à étudier en termes d'influence sur l'élaboration des stratégies d'oralisation. Ce processus permet d'obtenir une représentation de l'OT original qui prenne en compte l'ensemble des paramètres formels, repérés s'ils sont explicites ou déduits s'ils sont implicites.

REMARQUE. — Les deux processus (segmentation/extraction) s'appliquent alors récursivement jusqu'à ce que chaque objet textuel ne soit composé que d'OTs de type « phrases ponctuées sans marqueurs typo-dispositionnels » (ces derniers pouvant être qualifiés d'unités textuelles ou UT si l'on se réfère au MAT) ;

- phase de *reformulation* : il s'agit, lors de cette troisième étape, centrale à notre problématique, d'opérer un certain nombre de transformations qui visent à obtenir un

OT oralisable en fonction (1) des paramètres formels (déduits et repérés), (2) des relations de dépendance qu'ils entretiennent, et (3) du type de stratégie envisagée.

Moyennant une représentation finale adéquate, ces trois étapes permettraient, à partir d'un objet textuel écologique, de fournir en entrée d'un système de synthèse de parole, un objet « oralisable » et, le cas échéant, les schémas intonatifs associés à une stratégie particulière, sans modifications de l'architecture classique du TTS. Celui-ci pourra d'ailleurs s'affranchir de gérer ce que son système de prétraitement contrôle déjà, et prendre en compte, directement à travers les modules de phonétisation et de modélisation prosodiques, les nouvelles indications selon des spécifications qu'il nous impute de déterminer. Notre intérêt se porte ici sur le processus de reformulation : la mise en œuvre de ce processus nécessite (1) la conception d'un modèle des stratégies d'oralisation, (2) l'analyse des paramètres formels des OT étudiés, (3) l'élaboration de représentations informatiques pour les OT et chacune de leurs dérivations en fonction de la stratégie impliquée, (4) les spécifications pour la réalisation accoustico-prosodique par le TTS à partir de la représentation.

3.2. Le module de reformulation

Notre premier objectif pour aborder la phase de reformulation a été de concevoir un modèle de stratégies d'oralisation. L'axiome principal qui sous-tend le modèle postule la nécessité 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).

À l'instar de l'hypothèse d'équivalence fonctionnelle, proposée par le modèle d'architecture textuelle, entre la mise en forme matérielle de nature typodispositionnelle et la mise en forme matérielle de nature lexico-syntaxique, 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 prosodiques. 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 informationnel. La forme discursive d'un objet textuel particulier peut être reconstruite selon trois stratégies :

- par une description objective : forme discursive descriptive ;
- par un premier niveau d'interprétation, en attribuant un nom à la configuration : forme discursive dénomminative ;
- par une interprétation plus profonde des intentions du rédacteur qui se traduisent par l'utilisation de cet OT particulier : forme discursive interprétative.

Ces trois opérations pourront conduire ensuite à 3 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 ;

- 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 ou par un schéma prosodique spécifique, repéré comme pouvant être associé au type d'OT étudié indépendamment de tout contexte d'énonciation ;

- 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 termes de thème, topique, focus, rhème).

Une illustration des trois stratégies dans leur forme discursive peut être donnée à travers l'exemple suivant. Soit le titre d'un document :

L'ORALISATION DES TEXTES

S1 : « **Le texte en gros, centré, en gras et en capitale est** l'oralisation des textes »

S2 : « **Le titre de niveau n est** l'oralisation des textes »

S3 : « **Je vais vous parler maintenant de** l'oralisation des textes »

Les parties en gras seront les marques lexico-syntaxiques restituées à partir de la morpho-disposition du texte. Ce sont aussi celles qui pourront potentiellement être effacées au profit de marques sonores telles que décrites plus haut.

Nous avons choisi de détailler dans la section suivante l'opérationnalisation de la phase de reformulation pour l'OT que nous avons étudiée particulièrement en profondeur : les structures énumératives. Les stratégies décrites concernent les deux pôles de l'axe interprétatif du MORTELS : la forme discursive interprétative et la forme prosodique en contexte.

4. Opérationnalisation sur les structures énumératives

4.1. Paramètres formels des structures énumératives

Bien que ces structures soient très souvent utilisées, il n'existe pas de définition attestée sur l'objet énumération ; on appréhende souvent une énumération comme un procédé consistant à mettre sur le même plan d'équivalence (fonctionnelle et visuelle) des items. Cette vision nous semble trop restrictive au vu de la grande diversité des types de structures énumératives observés. En effet, des travaux sur l'énumération ont montré la diversité de ces structures : (Luc, 2000) propose une

typologie des énumérations non parallèles c'est-à-dire des énumérations ne respectant pas le parallélisme entre présentation et fonction des items. En opposition à ce terme, on parlera d'énumération parallèle pour les énumérations respectant ce parallélisme.

4.2. Construction de la forme discursive interprétative

La normalisation impose de s'être fixé un *pattern* prédéfini de métaphrases interprétatives susceptibles de recevoir la reformulation de n'importe quelle structure énumérative écologique.

L'inscription spatiale d'une structure énumérative parallèle de N items résulte de l'effacement d'une mise en forme matérielle de nature lexico-syntaxique dans une série de métaphrases formulables selon le métadiscours suivant :

L'auteur crée un texte identifié texte(1).

L'auteur distingue une énumération identifiée énumération(1).

L'auteur agence énumération(1) en N items identifiés [item(i) et]_{0<i<N} et item(N).

L'auteur chapeaute énumération(1) par un chapeau identifié chapeau(1).

L'auteur attache UT(1) à chapeau(1).

L'auteur compose chapeau(1) de UT(1).

[L'auteur attache UT(i+1) à item(i)]_{0<i<N}.

[L'auteur compose item(i) de UT(i+1)]_{0<i<N}.

L'auteur compose texte(1) de chapeau(1) [puis item(i)]_{0<i<N}

Dans l'objectif de produire un métadiscours le plus naturel possible pour constituer le texte oralisable nous ne retiendrons du métadiscours, dans le cadre de notre problématique particulière, que les éléments qui ne peuvent être rendus implicites. De plus, pour une raison d'intelligibilité plus grande du texte final nous avons choisi d'éliminer les différents identifiants et de les remplacer par une fusion des métaphrases qui rendent encore une fois implicite un certain nombre d'éléments du métadiscours original ; nous avons également paraphrasé pour cette même raison trois termes : *chapeauter* se traduira par *annoncer ci-après*, *items* par *points* et *attacher* par *préciser que* ; de plus, une information visuelle n'est pas prise en compte par ce métadiscours : le fait que le dernier item, lors d'une lecture de l'énumération, est en train d'être abordé. Cela entraîne une formulation particulière pour la dernière métaphore. Nous considérerons également que l'explicitation de l'acteur textuel est inutile car dans notre cas il s'agira toujours de l'auteur du texte. Ceci implique un certain nombre de refontes syntaxiques pour effacer « auteur » et l'acte performatif qui lui est rattaché. De plus, l'obligation d'utiliser un classifieur générique (« point ») est trop contraignante car l'oralisation des structures énumératives lorsque celui-ci est précisé de manière moins générique entraînerait une perte d'information. Nous nous imposons donc de repérer ou expliciter l'unité

de texte classifieur (que nous noterons UTC) des structures énumératives à oraliser (ceci implique de vérifier certains accords en genre).

Enfin, la gestion de l'exhaustivité et de l'ordonnancement des items (ou de l'absence d'indications pour statuer sur la valeur de ces paramètres) conduit à finalement retenir le *pattern* suivant (les parenthèses expriment la présence possible d'un des termes de la liste qu'elles bornent) :

(Les) N UTC (, parmi d'autres,) sont énuméré(e)s ci-après (dans l'ordre, dans le désordre). (Le, Un, La, Une) (premier, première) UTC (est, comprend, est de) UT(1). {(Le, Un, La, Une) 1^{ème} UTC (est, comprend, est de) UT(i)}_{1 ≤ i ≤ N} (Le, Un, La, Une) N^{ème} et (dernier, dernière) UTC (est, comprend, est de) UT(N).

4.3. Construction de la forme prosodique en contexte

L'objectif de cette phase de notre travail était de mettre en relation certaines marques lexico-syntaxiques des versions discursives dérivables à partir du *pattern* décrit précédemment, avec les configurations prosodiques qui pourraient s'y substituer. Pour cela, nous avons projeté sur les structures énumératives les intonèmes proposées par Rossi (1999) dans sa grammaire de l'intonation pragmatique du français. Les modèles prosodiques s'intéressant aux aspects pragmatiques du discours proposent la description prosodique de l'oral spontané. De ce fait la plupart des modèles considèrent que « l'intonation est toujours étudiée dans le cadre restreint de [...] la phrase. Appliquées aux énoncés d'une conversation, ces descriptions s'avèrent insuffisantes, voire inadéquates » (Mertens, 2001). Ainsi, leur intégration dans les TTS nécessiterait des modifications en profondeur des modules d'analyse syntaxique non fondée pour la conversation. Cependant le modèle de Rossi réduit cette incompatibilité entre modèles prosodiques de l'oral spontané et oralisation d'un écrit « réfléchi » puisqu'il considère au contraire que « les modèles construits pour décrire la structure syntaxique de la phrase et les outils conceptuels développés pour caractériser sa structure prosodique sont utilisables pour l'étude de la parole spontanée (Lacheret-Dujour, 1999 ; p. 193). Ainsi, les intonèmes décrits par ce modèle seraient utilisables directement « derrière » l'analyse syntaxique phrastique « classique » intégrée par les concepteurs de synthèse de parole à partir de texte. La projection de ce modèle sur l'OT « structure énumérative » nous a permis de supprimer un certain nombre de segments textuels dans notre forme discursive interprétative en les substituant par des configurations prosodiques impliquant les intonèmes de Rossi (Maurel, 2003).

Nous résumons, à travers un exemple, les différents processus et ressources impliqués dans notre système. Soit la structure énumérative :

Le Centre Social intervient entre autres dans :

- la gestion et le suivi de 6 appartements d'urgence que propose la commune ;

- l'organisation de réunions d'information collective ;
- les étapes du montage de projets de formation.

Celle-ci sera représentée de manière à respecter le schéma XML que nous avons défini pour les structures énumératives à partir de l'analyse de ses paramètres formels.

```
<SE type="parallèle" exhaust="non" ordon="neutre">
<AMORCE genre_classifieur= fem nom_classifieur= intervention >
    interventions du Centre Social
</AMORCE>
<ENUMERATION>
<ITEM cs= nom_sing > <NOYAU> la gestion et le suivi de 6 appartements d'urgence
</NOYAU> que propose la commune</ITEM>
<ITEM cs= nom_sing > <NOYAU> l'organisation de réunions d'information collective
</NOYAU> </ITEM>
<ITEM cs= nom_plur > <NOYAU> les étapes du montage des projets de formation </NOYAU>
</ITEM>
</ENUMERATION></SE>
```

Schéma qui est nécessaire et suffisant pour exploiter la technologie XSLT (*Extensible Stylesheet Language Transformation*) afin de produire la version discursive suivante :

Trois interventions du centre social, parmi d'autres, sont énumérées ci-après. Une première intervention est **la gestion et le suivi de 6 appartements d'urgence que propose la commune**. Une deuxième intervention est **l'organisation de réunions d'information collective**. Une troisième et dernière intervention comprend **les étapes du montage de projets de formation**.

La signification de la mise en forme utilisée dans ce nouveau texte est la suivante :

- le texte en caractère maigre romain est la partie prédéfinie du *pattern* ;
- le texte en gras est récupéré directement à partir du texte encadré par les balises AMORCE et ITEM ;
- le texte en italique est déduit de la valeur des attributs par recopie (*i.e.* reprise du classifieur) ou application de règles adéquates (*i.e.* gestion de l'exhaustivité, de l'ordonnancement ou de l'absence de prise de décision pour la valeur de ces paramètres) ;
- le texte souligné est obtenu par des opérations externes à la structure : comptage du nombre total de balises ITEM et du nombre de balises ITEM déjà prises en compte.

Nous considérons que ce type de texte peut être fourni à un logiciel de synthèse de la parole qui l'oralisera correctement en se basant sur une unité phrastique classique. Notre forme prosodique a ainsi la représentation de type SSML suivante :

```
<SE> <FOCUS type="étendu"> trois interventions du centre social </FOCUS>, parmi
d'autres, sont <BREAK strength ="strong"> <AUDIO src="uri"/> <PROSODY
pitch="calculé pour item 1/3"> <FOCUS type="rhématique"> la gestion et le suivi de
six appartements d'urgence </FOCUS> que propose la commune <BREAK
strength="medium"> </PROSODY> <PROSODY pitch=" calculé pour item 2/3">
<FOCUS type="rhématique"> l'organisation de réunions d'information collectives
</FOCUS> <BREAK strength =" medium "> </PROSODY> <PROSODY
range=Xst> <FOCUS type="rhématique"> les étape du montage des projets de
formation </FOCUS> </PROSODY> </SE>
```

L'implémentation a été réalisée en collaboration avec un partenaire industriel, Elan Speech, pour le moteur de synthèse *Proverbe Speech Engine*. Le logiciel de copie de la prosodie Prosel et un cycle de vie incrémental ont permis l'implémentation de nos spécifications SSML dans une version de leur moteur de synthèse. Tout d'abord, la structure prosodique de départ, sur laquelle s'est appuyée l'implémentation, est la synthèse du contenu articulable de la version prosodique en contexte, mais ponctuée d'une manière prédéfinie. Le fichier prosodique généré par la synthèse de parole à partir de ce texte est ensuite modifié par l'interprétation de notre marquage de type SSML puis réalisé accoustico-prosodiquement. Les fichiers prosodiques ainsi construits ne permettant pas de gérer le paramètre d'intensité, ce dernier n'a pas été implémenté. Le travail réalisé finalise donc pour un OT spécifique deux formes de l'axe interprétatif du MORTELS : la forme discursive interprétative (FDI) et la forme prosodique en contexte (FPC).

5. Conclusion et perspectives

Nous avons tout d'abord justifié, à travers le point de vue de typographes et de linguistes, l'intérêt de ces disciplines pour participer à répondre à notre problématique informatique de l'oralisation des structures visuelles des textes. L'approche proposée par le modèle d'architecture textuelle, pose un principe d'équivalence fonctionnelle entre différents types de marques de mise en forme matérielle (typographique, dispositionnelle et lexico-syntaxique) dont nous nous inspirons pour faire l'hypothèse d'une équivalence fonctionnelle entre des marques lexico-syntaxiques et des configurations prosodiques. Une analyse linguistique des objets textuels, mis en évidence par cette mise en forme matérielle, permet d'envisager l'opérationnalisation d'un premier jeu de stratégies de reformulation discursive.

Cette première partie nous a ainsi montré que, dans notre perspective d'automatisation de la transposition à l'oral de textes écrits pour être lus silencieusement, nous devons nous appuyer sur :

- une architecture différente des systèmes automatiques actuels puisque une unité textuelle de traitement telle que la phrase ponctuée est insuffisante pour le traitement de l'inscription spatiale du texte ; nous proposons une architecture dont l'unité textuelle de traitement serait l'objet textuel ; celui-ci subirait une série de transformations à travers différents modules (segmentation, extraction, reformulation) afin de fournir une version du texte dans une représentation interprétable à peu de coût par les logiciels de synthèse actuels ;

- des stratégies de reformulation reposant sur des transformations successives du texte ; nous proposons un modèle d'oralisation par reformulation des textes écrits pour être lus silencieusement qui a pour vocation de donner une vue exhaustive des stratégies envisageables et des relations qu'elles entretiennent entre elles ; ce modèle permet d'envisager la mise en œuvre du module de reformulation de notre architecture en fonction du type d'objet textuel susceptible d'être pris en compte et du rapport que l'on s'impose entre d'une part la complexité de mise en œuvre et d'autre part le naturel du résultat ;

- des typologies d'objets textuels pour définir quel type d'inscription spatiale les constituent et quels paramètres formels prendre en compte lors de l'oralisation.

Sur ces bases, nous avons opérationnalisé le module de reformulation pour un objet textuel particulièrement complexe : les structures énumératives. Se dégage ainsi une méthodologie générale, intéressante dans le cadre de l'ingénierie informatique, puisqu'elle nous permet de proposer :

- une formalisation de cet objet textuel sous la forme d'une DTD (*Document Type Definition*) que nous avons entièrement écrite ;

- des recommandations pour améliorer le langage de représentation SSML, chargé de représenter les aspects prosodiques d'un texte à oraliser, et ainsi obtenir une formalisation de la version prosodique sous la forme d'une DTD de type SSML ; cette dernière version de l'objet textuel est directement exploitable par un logiciel de synthèse de la parole à partir de texte à condition que nos spécifications prosodiques pour l'interprétation du marquage soient respectées (ces spécifications ont été implémentées pour la voix de synthèse *Proverbe d'Elan Speech*) ;

- l'utilisation de la norme XSLT pour dériver les différentes versions de l'objet textuel selon la stratégie utilisée (discursive ou prosodique).

Bien qu'il soit nécessaire de continuer le travail présenté « en largeur » sur d'autres objets textuels (titres, mise en gras, paragraphes...), nous travaillerons également sur une extension « en profondeur » (structures énumératives non parallèles).

Outre le cadre applicatif de la synthèse de parole à partir de texte, cette approche de l'inscription morphologique des textes permet de procéder à une analyse critique des normes de représentation émises par le World Wide Web Consortium (W3C),

que se soit au niveau des standards de formatage du texte (*i.e.* : XHTML, Cascading Style Sheet) ou au niveau de celles établies pour représenter les spécificités prosodiques de son oralisation (*i.e.*, *Speech Synthesis Markup Language*, *Aural Cascading Style Sheet*). Le cas échéant, nous proposons des nouvelles fonctionnalités, nécessaires à une représentation formelle des objets textuels qui permette d'envisager l'automatisation de l'étape de reformulation.

Enfin, nos objectifs sont également intégrer l'automatisation des phases de segmentation et d'extraction du système. Pour cela, nous travaillons actuellement à l'analyse automatique des documents électroniques, pour l'obtention de leur découpage en objets textuels dans leur représentation formelle définie par l'étude de l'étape de reformulation.

Remerciements

Ce travail a fait l'objet d'un soutien par le conseil régional de Midi-Pyrénées et le programme interdisciplinaire du CNRS « Société de l'information ».

6. Bibliographie

- Alessandro (D') C., Tzoukermann E., « Synthèse de la parole à partir du texte », *Traitement automatique des langues*, vol. 42, n° 1, Hermes Science Publications, Paris, 2001.
- Arabyan M., « Du nouveau sur le mythe des origines de la lecture silencieuse », in M. de Mattia et A. Joly eds., *De la syntaxe à la narratologie énonciative*, Gap-Paris, Ophrys, 2001.
- Boula de Mareuil P., Célérier P., Cesses T., Fabre S., Jobin C., Le Meur P.-Y., Obadia D., Soulage B., Toën J., Elan Text-To-Speech, « un système multilingue de synthèse de la parole à partir du texte », *Traitement automatique des langues*, vol. 42, n° 1, Synthèse de la parole à partir du texte, 2001, pp. 223-252.
- Chaytor H., *From script to print: an introduction to medieval literature*, Cambridge University Press, 1945.
- Cherbonnel B., Un module de pré-traitement multilingue de documents intégré au système de synthèse de parole à partir du texte du CENT, Document technique DT/267/LAA/TSS/RCP, CENT, Lannion, France, 1996.
- Duplan P., Jauneau R., *Maquette et mise en page*, Editions de l'Usine Nouvelle, Paris, 1986.
- Dutoit T., *An Introduction to Text-To-Speech Sythesis*, Kluwer Academic Publishers, 1997.
- Gardes Tamine J., *Pour une grammaire de l'écrit*, Belin, 2004.
- Harris Z., *Structures Mathématiques du Langage*, Dunod, 1971.
- Harris R., *La sémiologie de l'écriture*, Paris, Editions du CNRS, 1993.

- Luc C., Représentation et composition des structures visuelles et rhétoriques du texte, Thèse de doctorat, Université Paul Sabatier, 2000.
- Manguel A., *Une histoire de la lecture*, Acte Sud, 1998.
- Maurel F., Lemarié J., Vigouroux N., « Oralisation de structures visuelles : de la lexicosyntaxe à la prosodie », in *Actes de Interface Prosodique 2003 (IP2003)*, A. Mettouchi et G. Ferré (eds), 2003, pp. 137-142.
- Maurel F., « De l'écrit à l'oral : analyses et générations », in *Actes de 11e Conf. sur le Traitement Automatique des Langues Naturelles (TALN'2004)*, LPL-ENS/Fes-ATALA, ISBN 2-9518233-4-7, pp. 289-298, 2004.
- Maurel F., Transmodalité et multimodalité écrit/oral : modélisation, traitement automatique et évaluation de stratégies de présentation des structures « visuo-architecturales » des textes, Thèse de doctorat, Université Paul Sabatier, 2004.
- McLuhan M., *The Gutenberg galaxy*, London, Routledge & Kegan Paul, 1962.
- Mertens P., Goldman J.-P., Wehrli E., Gaudinat A., « La synthèse de l'intonation à partir de structures syntaxiques riches », *Traitement automatique des langues*, vol. 42, n° 1, Synthèse de la parole à partir du texte, 2001, pp. 145-192.
- Morel M., Lacheret-Dujour A., « “Kali”, synthèse vocale à partir du texte », *Traitement automatique des langues*, vol. 42, n° 1, Synthèse de la parole à partir du texte, 2001, pp. 193-221.
- Ong J., *Orality and literacy*, London, Methuen, 1982.
- Pascual E., “Integrating Text Formatting and Text Generation”, In : G. Adorni, M. Zock (Eds.) *Trends in Natural Language Generation, An Artificial Intelligence Perspective, 1993, Selected Papers. Lecture Notes in Computer Science 1036*, Springer, 1996, pp. 205-221.
- Rossi M., *L'Intonation, le Système du Français : Description et Modélisation*. Paris, Ophrys, 1999.
- Saenger P., “Silent reading: its impact on late medieval script and society”, *Viator*, 13, 1982, pp. 367-414.
- Sproat R., *Multilingual Text-To-Speech Synthesis : the Bell Labs Approach*, Kluwer Academic Publishers, 1998.
- Torzec N., Moudenc T., Emerard F., « Prétraitement et analyse linguistique dans le système de synthèse TTS CVOX », *Traitement automatique des langues*, vol. 42, n° 1, Synthèse de la parole à partir du texte, 2001, pp. 17-46.
- Virbel J., The Contribution of linguistic knowledge to the interpretation of text Structure, in J. André, V. Quint et R. Furuta (eds.), *Structured Documents*, Cambridge University Press, 1989, pp. 161-181.