MusiTAL : une partition à six mains pour le TAL

Marie Dozol[1] Paul Sabatier[2] Marie-Hélène Stéfanini[2]

(1) Aéropôle, D20H, route de l'aéroport, 13288 Marseille Cedex 9
(2) LIF, AMU, CNRS, 163 avenue de Luminy, 13288 Marseille Cedex 9

mdozol@webmail.alten.fr, paul.sabatier@lif.univ-mrs.fr, marie-helene.stefanini@lif.univ-mrs.fr

Résumé.

Nous présentons MusiTAL, une application d'analyse/synthèse de phrases dans le domaine de la musique, que nous avons conçue à partir des données du *Dictionnaire électronique des mots* (DEM) des Dubois et développée au moyen du logiciel ILLICO.

Abstract.

We describe a sentence analysis/synthesis application in music domain, MusiTAL, we have conceived from data described in Dubois' Electronic dictionary of words and developed by means of the ILLICO software.

Mots-clés : Dictionnaire électronique des mots des Dubois, DEM, ontologie, musique, analyse/synthèse de phrases, ILLICO.

Keywords: Dubois' electronic dictionary of words, ontology, DEM, music, sentences analysis/synthesis, ILLICO.

1 Introduction

Quel spécialiste de TAL n'a pas rêvé (ou rêve encore) de disposer de ressources linguistiques finement décrites dans un format approprié qui se prêteraient alors à une exploitation et à une intégration dans différents systèmes de TAL ? Les projets et initiatives ne manquent pas dans les communautés nationales et internationales, pour produire, développer, formater, enrichir et exploiter des ressources liées aux langues et à la faculté de langage (lexiques, grammaires, ontologies, etc.). Pour le français, les «TAListes » épris de la langue dans sa spécificité et ayant la volonté de formaliser ce qui peut l'être, n'ont pas manqué de regarder de près ce que pourraient leur apporter les travaux sur les lexiques-grammaires de Maurice Gross et de son équipe (Gross, 1994), de Maurice Salkoff (grammaire en chaîne) (Salkoff, 1973). Les ressources lexicales à grande couverture comme WordNet (Fellbaum, 1998), FrameNet (Baker, Fillmore, Lowe, 1998), VerbNet (Kipper-Schuler, 2005) ou Dicovalence (Van Den Eynde, Mertens, 2006) sont particulièrement utiles pour l'anglais. Ces ressources ont fait et font l'objet de formatages, de mises au point et d'exploitations dans la communauté TAL.

Dans cet article, nous nous intéressons à un autre ensemble de ressources développé par Jean Dubois et Françoise Dubois-Charlier. Nous présentons MusiTAL, un système d'analyse/synthèse de phrases dans le domaine de la musique, que nous avons conçu à partir des données que les Dubois ont décrites dans leur *Dictionnaire électronique des mots*.

2 Le Dictionnaire Electronique des mots (DEM)

Jean Dubois et Françoise Dubois-Charlier ont développé un dictionnaire électronique des mots (DEM) du français qui comprend 145333 entrées. Une présentation de DEM est donnée dans (Dubois, Dubois-Charlier, 2010), avec, à titre d'illustration, la description de 1 450 termes du domaine de la musique. Chaque entrée de DEM est constituée des rubriques suivantes :

- M : mot d'entrée (avec différenciation par des numéros en cas d'homonymie) ;
- CA : catégorie grammaticale (catégories traditionnelles complétées par une indication sur le référent (humain, chose, animal, masculin, singulier, invariable, etc.) ;
- GP : caractéristiques de formation pour le genre et le nombre (29 étiquettes pour la formation du féminin, 23 pour la formation du pluriel) ;
- DOM : indique le domaine ou "champ lexical/paradigmatique (186 domaines sont recensés), le niveau de langue (éventuellement), les régionalismes (francophonie : Belgique, Canada, Suisse) ;
- SENS : définition tirée des dictionnaires de référence (parfois ce peut être un synonyme).
Ex. : *chef d'orchestre*, SENS = "qui mène un orchestre" ;
- CONT (Contexte) : pour les adjectifs (ou adverbes), indique le nom (ou le verbe) prototype qu'il peut qualifier. Ex. : *antiphonique*, CONT = "chant", *moderato*, CONT = "jouer adv" ; pour les noms, complète la définition par un hyperonyme. Ex. : *crooner*, CONT = "chanteur" ; pour les verbes, indique un verbe prototype.
Ex. : *pianoter*, CONT = "N jouer" ;
- OP (Opérateur) : indique une sous-classe sémantique associée au mot en liaison avec CONT.
Ex. : *chef d'orchestre*, OP = "spé" pour spécialité ;
- OP1 (Classe de verbe associée) : pour les noms, adjectifs et adverbes, indique la classe de verbes avec lesquels ils peuvent se combiner. Il s'agit des 14 classes sémantiques génériques de verbes, sous-catégorisées en 54 classes sémantico-syntaxiques (selon les oppositions être vivant/non-animé et propre/figuré (ou métaphorique)) qui se répartissent en 248 sous-classes syntaxiques selon leurs constructions syntaxiques et leur paradigme lexical.

3 MusiTAL = DEM (Musique) + ILLICO + GNF

Dans le cadre de l'initiative FondamenTAL[1], nous nous sommes intéressés à concevoir une application à partir d'un sous-ensemble des mots de DEM, à savoir celui constitué par les noms, adjectifs, verbes et adverbes du domaine de la musique, soit près de 1 450 entrées de DEM. L'application développée permet d'analyser, de synthétiser (ou « générer ») ou d'aider à composer des phrases dans le domaine de la musique, comme par exemple :

Les clochettes tintinnabulent. La guitare de Max est désaccordée. Luc entonne l'Internationale. Léa joue du saxophone. Le balafon est un idiophone à percussion. Marie a l'oreille musicale. Léo siffle comme un merle. Quelles sont les cantates composées par Bach ?

L'application a été développée au moyen du logiciel ILLICO (Pasero, Sabatier, 2008). Les phrases sont analysées/synthétisées/composées à partir de GNF, une grammaire noyau décrivant les constructions fondamentales du français. Une représentation sémantique de type logique est automatiquement associée à chaque phrase bien formée.

Par exemple, pour la phrase : *Le chef de chœur chante comme une casserole*.

MusiTAL produit la représentation sémantique (Figure 1) :

[1] FondamenTAL : http://www.talep.lif.univ-mrs.fr/FondamenTAL.html

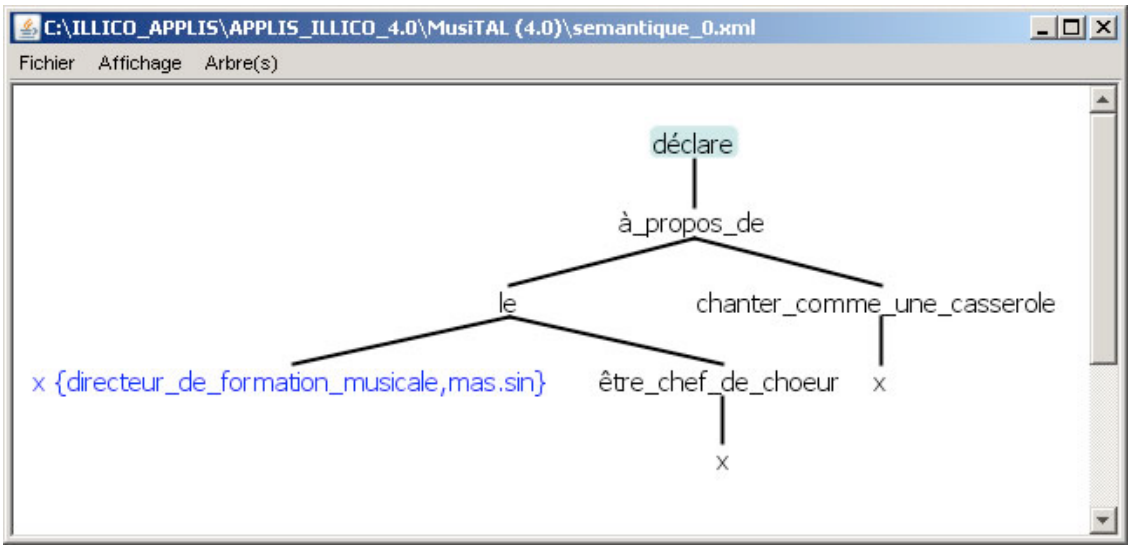

Figure 1 – MusiTAL : représentation sémantique

L'important dans ce type d'application est de pouvoir dire si une phrase analysée est bien formée. Si c'est le cas, une (ou plusieurs, en cas d'ambiguïté) représentation sémantique est automatiquement associée. Si la phrase est mal formée, des corrections lexicales, syntaxiques et conceptuelles sont proposées. Les phrases synthétisées doivent être bien sûr bien formées. Ces contraintes en analyse et en synthèse nécessitent une description très fine des données linguistiques, et cela à différents niveaux de bonne formation. La définition et la formalisation des contraintes de bonne formation lexico-morpho-syntaxiques ne constituent pas une tâche insurmontable. La littérature abonde de descriptions et de règles formelles pour ces domaines. C'est plutôt au niveau de la bonne formation conceptuelle que la tâche à réaliser est importante. Par exemple, pour le domaine qui nous intéresse ici, il faut pouvoir considérer que, par exemple, les phrases suivantes sont plutôt perçues comme conceptuellement malformées :

La guitare tintinnabule. Max accorde le triangle. Luc remplace une corde de la clarinette. Max souffle dans le sistre.

Une phrase est conceptuellement bien formée si la représentation sémantique associée décrit une situation conceptuellement possible, c'est-à-dire, de façon plus formelle, si les relations et les individus qu'elle met en jeu sont compatibles. L'expression d'une telle compatibilité peut être formulée au niveau du lexique et des règles syntaxiques au moyen de « traits sémantiques » spécifiques. Dans ILLICO, cette compatibilité peut être formulée de façon plus modulaire et déclarative au moyen de ce qu'on appelle le modèle conceptuel. Le modèle conceptuel rend compte de phénomènes relevant du domaine traditionnel de la sémantique dite lexicale. Le caractère conceptuellement bien formé d'une phrase est établi dans ILLICO à partir de deux types de contraintes conceptuelles : les contraintes de domaines et les contraintes de connectivité. De façon pratique pour ce qui est des contraintes de domaines, la vérification du caractère conceptuellement bien formé d'un énoncé consiste simplement à vérifier la compatibilité des types associés aux individus, aux relations et aux fonctions des éléments de la représentation sémantique intermédiaire de l'énoncé. Cela suppose que les constantes, les variables logiques et les symboles relationnels et fonctionnels soient typés conceptuellement. Le modèle conceptuel contient les connaissances permettant d'associer un type conceptuel aux constantes, aux variables logiques et aux symboles relationnels et fonctionnels de la représentation intermédiaire. Le traitement conceptuel consiste alors à vérifier leur compatibilité (Pasero, Sabatier, 2008).

Pour la mise au point des contraintes conceptuelles de domaine, les dictionnaires des Dubois trouvent tout leur intérêt. En effet, comme pour le LVF (Dubois, Dubois-Charlier, 1997), l'intérêt de DEM réside en particulier dans la nature des informations sémantiques qu'il contient, avec pour chaque entrée les trois rubriques CONT (Contexte), OP (Opérateur) et OP1(Classe de verbe associée).

Exemple : Accordéoniste CONT = "N qui joue de", OP = "spéc", OP1 = "C1c3" signifie :

un accordéoniste est une personne qui joue d'un instrument (défini dans la rubrique SENS), dont c'est la spécialité ("spéc"), ce qui en fait le sujet de verbes exprimant l'idée d' "émettre des sons à fonction expressive et esthétique" (C1c3 est une des 54 classes sémantico-syntaxiques typant les 25 609 emplois de verbes dans LVF). Pour une présentation détaillée des classes du domaine de la musique, on consultera (Dubois, Dubois-Charlier 2010).

La figure 2 donne un extrait du modèle conceptuel que nous avons construit à partir des indications fournies dans le DEM, pour ce qui concerne la classification et la hiérarchie des différents domaines conceptuels (ou "classes" ou "types") associés aux noms.

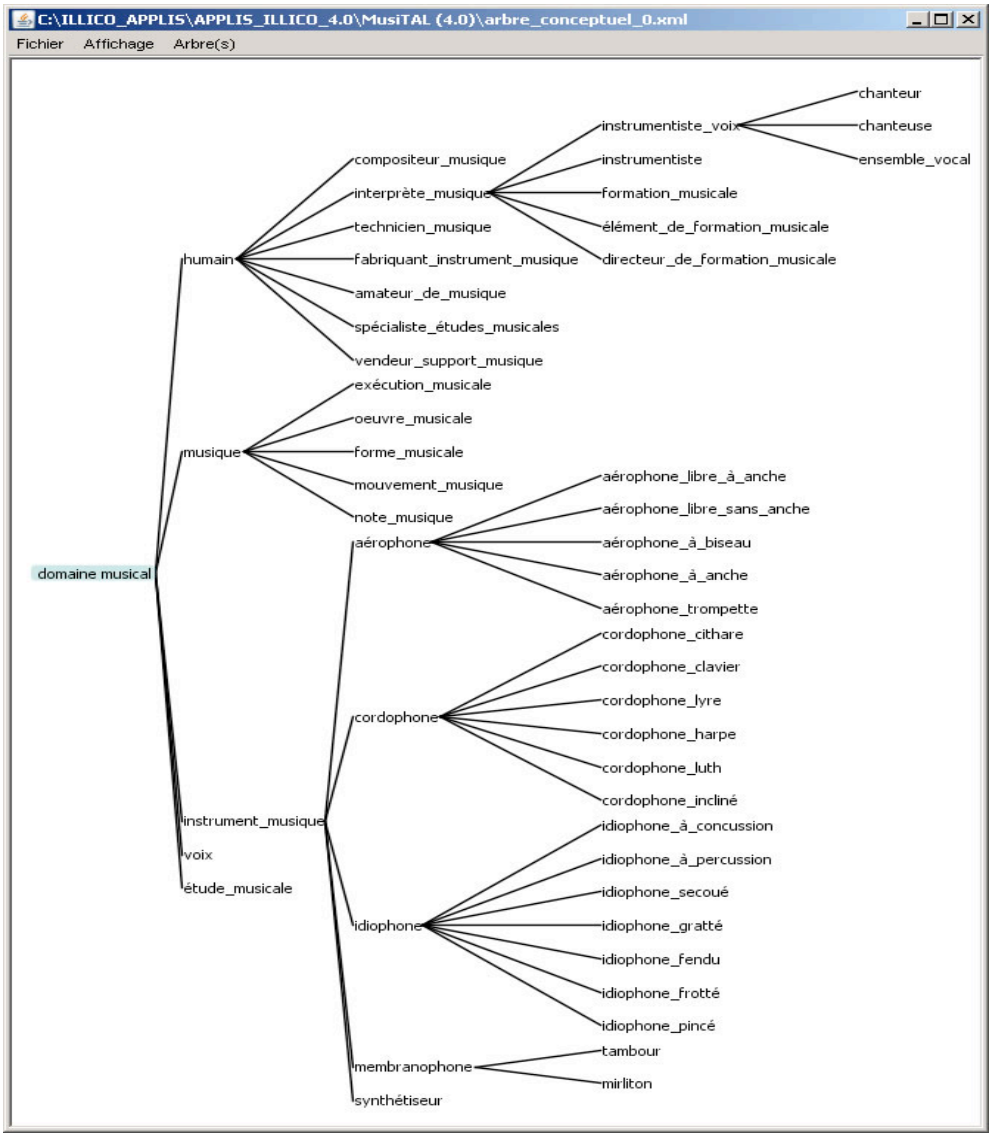

Figure 2 – MusiTAL : extrait du modèle conceptuel

Les feuilles de l'arbre conceptuel correspondent aux noms. Par exemple, le type "formation_musicale" est associé aux noms suivants : *bagad, big band, boeuf, chanterie, chantrerie, clique, cobla, combo, ensemble, fanfare, formation, gamelan, harmonie, jam-session, jazz-band, manécanterie, musique militaire, octuor, orchestre musette, orchestre, orphéon, otteto, philharmonie, quartet, quartette, quatuor, quintet, quintette, ripieno, septuor, sextuor, takht, taraf, trio, tutti*.

Les actants (sujet, objet) des verbes reçoivent un type conceptuel. Par exemple, pour le verbe *diriger*, nous avons, entre autres relations, la relation suivante : *diriger (humain, formation musicale)*

La mise au point de MusiTAL a tiré profit des fonctionnalités offertes par la version 4.0 d'ILLICO, en particulier celles qui permettent de formuler différents types de contraintes sur les expressions (mots, syntagmes, propositions, phrases, etc.) à analyser ou à synthétiser. On peut tester et évaluer les compétences et performances linguistiques et cognitives de systèmes de TAL en leur soumettant des expressions à analyser. Une autre manière de procéder est de demander à ces systèmes de produire des expressions vérifiant un ensemble de contraintes précises et de vérifier ensuite si l'ensemble des expressions produites est celui attendu. ILLICO offre la possibilité de formuler de façon modulaire et dynamique différents types de contraintes sur les expressions, comme par exemple des contraintes sur les niveaux de bonne

formation (lexical, syntaxique, conceptuel et contextuel), des contraintes sur la structure des expressions (formulées au moyen de coupes syntaxiques totales ou partielles), des contraintes lexicales (mots autorisés ou interdits), ou des contraintes sur la longueur des expressions.

4 Conclusion

L'application MusiTAL, que nous avons conçue et développée, nous a permis de mesurer la qualité des ressources linguistiques développées par F. Dubois et J. Dubois et leur intérêt pour le TAL[2]. L'apport du DEM, comme celui de LVF résident dans la finesse des descriptions sémantiques et conceptuelles systématiquement associées aux entrées de leurs dictionnaires. On peut alors penser que le recours à des ressources qui feront le lien entre celles développées par les Dubois et celles issues des autres travaux fondamentaux[3] mentionnés dans l'introduction se révèlera hautement bénéfique pour améliorer la qualité des systèmes de TAL.

Remerciements

Nous tenons à remercier Françoise Dubois-Charlier et Jean Dubois pour les échanges que nous avons eus et pour la qualité des ressources qu'ils ont développées et mises à notre disposition.

Références

DUBOIS, J., DUBOIS-CHARLIER, F. (1997). *Les verbes français*, Larousse-Bordas.

DUBOIS, J., DUBOIS-CHARLIER, F. (2010). La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration, *Langages*, 179-180, 31-56.

FELLBAUM, C. (2010). *WordNet* : *An Electronic Lexical Database*, Cambridge (MA). MIT Press.

FILLMORE C.J., LOWE J. B. (1998). The Berkeley FrameNet project. *COLING-ACL'98*, 86-90.

GROSS, M. (1994). Constructing Lexicon-grammars, Computational Approaches to the Lexicon. Atkins and Zampolli (eds.), Oxford Univ. Press, 213-263.

HADOUCHE F., LAPALME G. (2010). Une version électronique du LVF comparée à d'autres ressources lexicales, *Langages*, 179-180, 193-220.

KIPPER-SCHULER1 K. (2005). *VerbNet : A broad-coverage, comprehensive verb lexicon*. PhD Thesis, University of Pennsylvania.

LEEMAN, D., SABATIER, P., DIR. (2010). Empirie, Théorie, Exploitation : l'exemple du travail de Jean Dubois sur les verbes français, *Langages*, 179-180.

PASERO, R., SABATIER, P. (2008). ILLICO : Principes, connaissances et formalismes & Guide d'utilisation, Document Web, LIF.

PASERO, R., SABATIER, P. (2008). GNF : Une grammaire noyau du français, Document Web, LIF.

SALKOFF, M. (1973). *Une grammaire en chaîne du français Analyse distributionnelle,* Dunod.

VAN DEN EYNDE K., MERTENS P. (2006). Le dictionnaire de valence Dicovalence : *Manuel d'utilisation*, Leuven : Université de Leuven. [http://bach.arts.kuleuven.be/dicovalence/manuel_061117.pdf].

[2] Pour d'autres exemples d'exploitation de ces ressources, voir le numéro de la revue *Langages* (Leeman, Sabatier, 2011).

[3] Une étude comparative du LVF avec différentes ressources lexicales a été définie par (Hadouche, Lapalme, 2010).