*Studii de lingvistică 3, 2013, 187 - 204*

# Aide à la rédaction: le système TAEMA. De l'intérêt d'exploiter des ressources lexicales en linguistique informatique

TAEMA system: the advantages of exploiting lexical resources in NLP

Pierre-André Buvet[1]

**Abstract:** We discuss here the need to implement resources which are exhaustive and enriched by strong metalinguistic descriptions to develop systems which deal with textual information. Firstly, we present the data model used to create these resources; secondly, we specify the nature of these resources; and finally we show how they allow the TAEMA system to work. The purpose of this system is to help learners of French as a second language to diversify their expression with regard to the vocabulary of affects.

**Key words:** system of assistance with writing, linguistic resources, predicates, affects.

## 1. Introduction

TAEMA signifie « Traitement Automatique pour l'Ecriture de Mots Affectifs ». Il s'agit d'un prototype de Système d'Aide à la Rédaction (SAR) dont la finalité est d'aider les apprenants français langue seconde à diversifier leur expression pour ce qui concerne le vocabulaire des affects (cf. Buvet & Issac 2006)[2]. La qualité des résultats proposés par le prototype est imputable à celle des ressources linguistiques qu'il exploite. Nous discutons ici de la nécessité d'implémenter des ressources exhaustives et enrichies de descriptions métalinguistiques de qualité pour développer des systèmes qui traitent efficacement l'information textuelle. Après avoir présenté le modèle de données que nous utilisons pour créer ces ressources, nous précisons quelle est leur nature puis nous montrons comment elles permettent le bon fonctionnement de TAEMA.

[1] CNRS –Université Paris 13-Sorbonne Paris Cité, Laboratoire LDI, UMR 7187; pabuvet@ldi.univ-paris13.fr.

[2] TAEMA est le fruit d'un travail collectif impliquant, outre l'auteur du présent article, Fabrice Issac du laboratoire LDI de l'Université Paris 13; Aurélie Joseph, ingénieure de la société ITSOFT, et la première promotion du master pro TILDE de l'Université Paris 13.

## 2. Contexte méthodologique

La linguistique informatique ne se contente pas d'appliquer des outils informatiques aux documents pour traiter leur contenu textuel, elle implique également le recours à un cadre théorique suffisamment formel pour créer des ressources linguistiques. Ces ressources sont implémentées dans des systèmes dédiés au traitement de l'information textuelle afin que les dits outils les utilisent. Nous précisons dans cette section le cadre théorique de référence, le concept de prédicat d'affect et la définition de la sémantique qui résulte du modèle de données utilisé.

### 2.1. Modèle de données

La théorie des trois fonctions primaires est le modèle de données auquel nous nous référons. Elle s'inscrit dans la lignée des théories qui ont comme point de départ les analyses linguistiques de Zellig S. Harris. Sa finalité est d'expliquer les mécanismes langagiers en privilégiant le lexique comme objet d'étude. Il s'agit d'analyser conjointement les propriétés morphologiques, syntaxiques et sémantiques des unités linguistiques selon qu'elles fonctionnent comme des prédicats (fonction prédicative), des arguments (fonction argumentale) ou des actualisateurs (fonction actualisatrice).

Le modèle de données distingue trois niveaux de fonctionnement: le niveau logico-sémantique, le niveau énonciatif et le niveau interprétatif (cf. Buvet 2011). Les premier et deuxième niveaux sont impliqués dans l'émission de l'information, les premier et troisième niveaux le sont dans la réception de l'information. Le second est surtout centré sur le locuteur tandis que le troisième est plutôt focalisé sur l'interlocuteur. Les discours des locuteurs sont les seules données observables, ils relèvent du seul niveau énonciatif. Les deux autres niveaux procèdent d'une modélisation fondée sur l'étude de l'articulation entre le premier niveau et le second, d'une part, le second niveau et le troisième, d'autre part.

La question du savoir linguistique est transversale aux trois niveaux de fonctionnement, la transmission et la réception de l'information présupposant la maîtrise d'un code commun (cf. Mattelart A. & Mattelart M. 2004). Le dialogue implique la maîtrise de l'encodage et du décodage de l'information par tous les acteurs de la communication, (cf. Rosier 2008). La réversibilité des statuts de locuteur et d'interlocuteur dans le cadre d'échange d'informations est illustrée par la catégorie de la personne et de ses marqueurs spécifiques (cf. Charaudeau 1992). Malgré le caractère transversal du savoir linguistique, la transmission d'une information est unidirectionnelle. C'est pourquoi les schémas de l'information,

quels qu'ils soient, sont orientés du locuteur vers l'interlocuteur (cf. Shannon 1948). Dans la théorie des trois fonctions primaires, les structures prédicat-argument sont le point de départ du modèle de données. Elles relèvent du niveau logico-sémantique. L'ensemble des structures prédicat-argument correspond à une composante du savoir linguistique, celle qui concerne la représentation langagière du monde. Cette représentation est sous-tendue par une conception du monde en termes d'entités simples et complexes et de relations entre ces entités dont les formes linguistiques sont les structures prédicat-argument. Les langues parlées, quelles qu'elles soient, possèdent un ensemble de structures prédicat-argument qui sont propres à chacune d'entre elles dans la mesure où elles ne partagent pas les mêmes formes linguistiques.

Les structures prédicat-argument sont donc conçues comme autant d'éléments fondamentaux d'une composante d'un savoir linguistique partagé. Elles permettent de formuler des contenus propositionnels et sont instanciés dans les discours par le biais de l'actualisation, dont le rôle est de produire des énoncés bien formés relativement à des situations d'énonciation particulières. L'actualisation fait appel aux catégories énonciatives telles que l'aspect, la modalité, la personne, le temps, etc. Ces catégories permettent l'ancrage des structures prédicat-argument dans un discours donné en fonction de la position du locuteur par rapport à ce qu'il énonce. L'actualisation est supportée par les actualisateurs qui, tant du point de vue de leur forme que de celui de leur combinatoire avec les prédicats et les arguments, constituent une autre composante du savoir linguistique partagé, la grammaire d'une langue.

Le concept d'emploi prédicatif est central pour expliquer l'instanciation d'une structure prédicat-argument dans un discours du fait de l'actualisation. Ce concept relève du niveau énonciatif et, à ce titre, il procède d'observables. L'analyse d'un emploi prédicatif est fondée sur deux sortes de propriétés: des propriétés de nature sémantique et des propriétés formelles (cf. Buvet 2009a). Le mode de fonctionnement d'un emploi prédicatif est expliqué en corrélant des propriétés formelles aux propriétés sémantiques. Ces dernières sont au nombre de quatre: la racine prédicative, la classe sémantique, le type sémantique et l'aspect inhérent. Les autres sont au nombre de trois: la construction, la distribution morphosyntaxique et la distribution sémantique. Un emploi prédicatif peut avoir des propriétés communes avec d'autres emplois prédicatifs, en l'occurrence sa racine prédicative et sa classe sémantique, dans la mesure où ces emplois résultent, au niveau logico-sémantique, d'une structure prédicat-argument commune. Cette dernière est définie par une racine prédicative et une classe sémantique, autrement dit ces propriétés relèvent de deux niveaux de fonctionnement: le niveau logico-sémantique et

le niveau énonciatif. C'est l'observation et l'analyse du seul niveau énonciatif qui permet d'inférer quelles sont les propriétés sémantiques des structures prédicat-argument au niveau logico-sémantique. De même, la distribution sémantique observée au niveau énonciatif est en partie révélatrice du domaine d'arguments d'un prédicat au niveau logico-sémantique. Pour autant, des emplois prédicatifs issus d'une même structure prédicat-argument ne partagent pas nécessairement la même distribution car, selon les emplois prédicatifs, tous les arguments ne sont pas instanciés et lorsqu'ils le sont, ils n'occupent pas nécessairement les mêmes positions (cf. *infra*). Le type sémantique et l'aspect inhérent ainsi que la construction et la distribution morphosyntaxique sont des propriétés qui procèdent uniquement du niveau énonciatif.

Pour illustrer le concept d'emploi prédicatif, nous examinons les emplois verbaux, adjectivaux et nominaux suivants: *accompagner* dans (1a) *Il* ***accompagne*** *la viande de légumes cuits à l'eau* et dans (2a) *Le père* ***accompagne*** *son fils jusqu'à la porte de l'école; accompagné* dans (1b) *La viande est* ***accompagnée*** *de légumes cuits* et (2b) *Le fils est* ***accompagné*** *de son père; accompagnement* dans (1c) *Il a fait un* ***accompagnement*** *de légumes cuits avec la viande*. En premier lieu, précisons qu'il y a d'autres racines prédicatives accompagn- correspondant à des structures prédicat-argument différentes[3]. Ces structures sont constituées de prédicats formellement identiques mais dont les domaines d'arguments sont distincts. Pour ce qui est des exemples, il y a deux prédicats triadiques, c'est-à-dire des prédicats dont les domaines d'arguments sont ternaires (cf. Harris 1976). Pour autant, ces domaines ne sont pas semblables:

- accompagn-1 (HUMAIN, ALIMENT, ALIMENT)
- accompagn-2 (HUMAIN, HUMAIN, LOCATIF)

Les structures ont des modes d'instanciation distincts: le prédicat accompagn-1 est instancié sous trois formes, une forme verbale dans (1a), une forme adjectivale dans (1b) et une forme nominale dans (1c), le prédicat accompagn-2 l'est sous deux formes, une forme verbale dans (2a) et une forme adjectivale dans (2b). L'emploi verbal et l'emploi adjectival du prédicat accompagn-1 ont des propriétés sémantiques différentes de celles de l'emploi verbal et de l'emploi adjectival du prédicat accompagn-2 du fait qu'ils n'ont pas les mêmes propriétés distributionnelles.

Dans (1a), le verbe *accompagner* a les propriétés suivantes[4]:

[3] Il y a d'autres racines prédicatives accompagn-; par exemple, celle en rapport avec cet emploi d'*accompagner*: *Il* ***accompagne*** *son patient dans sa rémission*.

[4] L'absence du complément second donne lieu à un autre emploi de accompagner (*Le père* ***accompagne*** *son fils*) tel que l'aspect inhérent est le duratif inaccompli.

propriétés sémantiques[5]

- racine prédicative: accompagn-1
- classe sémantique: GARNITURE
- type sémantique: action
- aspect inhérent: duratif accompli

propriétés formelles

- construction: X0 V X1 PREP2 X2 (PREP2 = de+avec)
- distribution morphosyntaxique: X0= GN/X1=GN/X2=GN
- distribution sémantique: X0=HUMAIN/X1=ALIMENT/ X2=ALIMENT

Dans (2a) il a les propriétés suivantes:

propriétés sémantiques

- racine prédicative: accompagn-2
- classe sémantique: ESCORTE
- type sémantique: action
- aspect inhérent: duratif inaccompli

propriétés formelles

- construction: X0 V X1 PREP2 X2 telle que PREP2 = à+jusqu'à
- distribution morphosyntaxique: X0= GN/X1=GN/X2=GN
- distribution sémantique: X0=HUMAIN/X1=HUMAIN/ X2=LOCATIF

Dans (1b), l'adjectif *accompagné* a les propriétés suivantes:

propriétés sémantiques

- racine prédicative: accompagn-1
- classe sémantique: GARNITURE
- type sémantique: état
- aspect inhérent: résultatif[6]

propriétés formelles

- construction: X0 être A PREP1 X1 telle que PREP1 = de

[5] Les différentes propriétés sémantiques sont mises en évidence avec des critères formels (Buvet & Grezka 2007).
[6] Sur l'état résultatif, cf. Cresseils 2000.

- distribution morphosyntaxique: X0= GN/X1=GN
- distribution sémantique: X0=ALIMENT/X1=ALIMENT

Dans (2b), il a les propriétés suivantes:

propriétés sémantiques

- racine prédicative: accompagn-2
- classe sémantique: ESCORTE
- type sémantique: état
- aspect inhérent: provisoire

propriétés formelles

- construction: X0 être A PREP1 X1 telle que PREP1 = de+par
- distribution morphosyntaxique: X0= GN/X1=GN
- distribution sémantique: X0=HUMAIN/X1=HUMAIN

Dans (1c), le substantif *accompagnement* a les propriétés suivantes:

propriétés sémantiques

- racine prédicative: accompagn-1
- classe sémantique: GARNITURE
- type sémantique: action
- aspect inhérent: duratif accompli

propriétés formelles

- construction: X0 Vsupport N PREP1 X1 (PREP2 X2) telle que Vsupport = faire PREP1 = de PREP2=avec
- distribution morphosyntaxique: X0= GN/X1=GN
- distribution sémantique: X0=HUMAIN/X1=ALIMENT/ X2=ALIMENT

### 2.2. Les prédicats d'affect

Les prédicats d'affect se rapportent à l'intériorité de l'homme du point de vue de ce qu'il ressent psychologiquement mais aussi, de façon parallèle, physiologiquement et cognitivement (cf. André 2009, Buvet *et al.* 2005, Cyrulnik 1991, Tutin *et al.* 2006). En tant que prédicats, ils correspondent à des relations telles que leur nature affective implique d'avoir une entité humaine comme point d'ancrage de la relation. Les domaines d'arguments des prédicats d'affect sont unaires (par exemple *maussade*) ou binaires (par exemple *haïr*) de telle

sorte que les relations ont un caractère réflexif dans le premier cas de figure et sont orientées dans le second cas de figure. Lorsque le prédicat est dyadique, l'orientation s'effectue vers l'entité humaine; cette dernière est donc à la fois le point d'ancrage et le point d'arrivée de la relation. La notion d'expérienceur rend compte de cette double nature de l'individu affecté (cf. Anscombre 1995, 1996). L'autre argument concerne le point de départ de la relation et, en tant que tel, il s'interprète toujours comme une cause (cf. Van de Velde 1995). Sa nature est non contrainte et varie selon les prédicats d'affect.

Un prédicat monadique, par définition, ne comporte aucune cause dans son domaine d'arguments dans la mesure où une cause implique une autre entité qu'elle détermine et explique (cf. Nazarenko 2000). L'argument étant limité à une entité humaine, celle-ci a trois propriétés: elle est le point d'ancrage d'un affect spécifique, le point de départ et le point d'arrivée de la relation, ce qui explique son caractère réflexif.

La dénomination des affects a donné lieu en français a un lexique varié qui s'avère parfois très imagé, par exemple, *avoir les boules* comme équivalent d'*éprouver de la contrariété*. Les conditions d'occurrences des unités lexicales sont déterminées par les emplois prédicatifs qu'elles impliquent (cf. *supra*). Le lexique relatif à des affects a deux caractéristiques essentielles: (i) la polymorphie de la plupart des unités monolexicales (*aimer*, *amour* et *amoureux* dans *Il l'aime, Il éprouve de l'amour pour elle* et *Il est amoureux d'elle* sont trois formes correspondant à un même prédicat; il en est de même pour *énervement*, *énervant*, et *énerver* dans *Il ressent un certain énervement*, *Il est énervant*, *Il l'énerve*); (ii) le figement des unités polylexicales (*Il prend facilement la mouche*)[7].

La liste des prédicats d'affect est constituée d'environ 270 verbes, 250 noms, 450 adjectifs, qui correspondent à près de 150 racines prédicatives. Une racine prédicative est la base commune des différents emplois d'un même prédicat. Par exemple, le prédicat verbal *idolâtrer*, le prédicat ajdectival *idolâtre* et le prédicat nominal *idolâtrie* partagent la même racine prédicative *idolâtr-*. Les prédicats d'affect sont définis par une soixantaine de classes sémantiques. Les classes sémantiques sont fondées sur la base de relations de synonymie entre les emplois prédicatifs. La contiguïté sémantique des prédicats d'affect a donné lieu, entre autres, aux classes suivantes: AMITIE, AMOUR, ANTIPATHIE, HAINE, JALOUSIE, PITIE, RESPECT, RESSENTIMENT, SYMPATHIE.

Une étude fine des prédicats d'affect permet de les sous-catégoriser en fonction des propriétés sémantiques et syntaxiques de leurs emplois. Les sous-catégories les plus remarquables sont les prédicats d'émotion, les prédicats d'humeur et les prédicats

[7] Pour ce qui est du dernier point, divers tests transformationnels permettent de le constater (cf. Gross 1996b et Mejri 1997).

de sentiment. Par exemple, la première sous-catégorie concerne des prédicats décrits par les classes sémantiques COLERE, ENTHOUSIASME, JOIE, PEUR, TRISTESSE.

Les prédicats de sentiment sont tous diadyques; ils impliquent une relation orientée entre deux individus telle que l'être humain qui n'est pas le siège de l'affect en est la cause (cf. Van de Velde 1995). Les structures argumentales de leurs emplois attestent souvent du caractère binaire de leurs domaines d'arguments: *Il le désteste; Il l'adore*. Les emplois dont la forme est un ajdectif monovalent ont tous comme sujet l'être humain qui succite le sentiment: *Il est détestable, Il est adorable*. Les prédicats d'humeur sont tous monadiques. Leurs emplois sont essentiellement des adjectifs et des noms tels que leur structure argumentale comporte uniquement l'être humain affecté: *Il est maussade; sa maussaderie*. Les prédicats d'émotion sont diadyques mais les structures argumentales des emplois prédicatifs ne comportent pas nécessairement les arguments sous-catégorisés par les prédicats: *Cela effraie les enfants*; *Cela est effrayant*; *Il ressent un certain effroi (E + à la vue de cela)*.

Les prédicats d'humeur ne sont jamais instanciés dans des groupes prépositionnels (cf. Melis 2003). Seuls les substantifs correspondant à des prédicats d'émotion et de sentiment sont autorisés en position **N** dans la construction **PREP (DET + E) N**, mais les occurrences des substantifs dépendent souvent de la nature des prépositions: *avec amour / avec joie / *avec maussaderie; par amour / *par joie + *par maussaderie; *d'amour / de joie / *de maussaderie / *à son plus grand amour / à sa plus grande joie / à sa plus grande maussaderie*. Les prédicats d'émotion, contrairement aux prédicats d'humeur et de sentiment, ont des emplois nominaux qui sont compatibles avec des verbes de MOUVEMENT_PHYSIOLOGIQUE ou de MANIFESTATION_PHYSIOLOGIQUE: *Il tremble de (peur + *haine + *maussaderie); Il frissonne de (peur + *haine + *maussaderie)*; *Il dégouline de (peur + *haine + *maussaderie); Il suinte (la peur + *la haine + la maussaderie*).

### 2.3. De la sémantique lexicale à la sémantique énonciative

L'analyse des prédicats présentée jusqu'à présent relève essentiellement de la sémantique lexicale. L'affect étant une composante fondamentale de la subjectivé (cf. James 1894), cette dernière, en tant que notion linguistique, est indissociable de la sémantique énonciative (cf. Benveniste 1966). Il s'ensuit que les prédicats d'affect constituent un observatoire privilégié pour étudier l'articulation entre la sémantique lexicale et la sémantique énonciative.

La sémantique occupe une place centrale pour étudier les faits de langue. Sa conception diffère selon que l'analyse porte sur des unités lexicales ou sur des unités énonciatives. Les deux types

de traitement sont souvent présentés comme antagonistes dans la mesure où le premier procéderait de la linguistique de la phrase, le second de la linguistique de l'énoncé (cf. Reboul & Moeschler 1998). Nous proposons de dépasser ce clivage en tenant compte de tous les facteurs sémantiques, c'est-à-dire aussi bien d'ordre lexico-syntaxique que d'ordre énonciatif, et de déterminer en quoi leur combinaison contribue à l'interprétation ciblée des textes (cf. Buvet à paraître).

Nous réfutons l'opposition entre sémantique lexicale et sémantique textuelle et nous appelons linguistique du discours notre approche de la sémantique. La linguistique du discours se caractérise par une approche intégrée telle que les différentes facettes de l'analyse sémantique sont considérées comme se rapportant conjointement au lexique, à l'énonciation et à la compréhension. Il ne s'agit pas d'une accumulation de traitements sémantiques composites mais d'un traitement homogène qui prend appui sur le lexique pour analyser les autres phénomènes langagiers. L'objectif est de s'appuyer sur le lexique et la sémantique lexicale pour développer une linguistique du discours exhaustive. La méthode est de procéder avec l'approche intégrée. L'évaluation s'effectue avec des applications dédiées au traitement de l'information.

Les prédicats d'affect sont révélateurs de la continuité entre sémantique lexicale et sémantique énonciative car, lorsque le discours argumentatif cherche à persuader, il permet d'exprimer toutes sortes d'états affectifs afin d'impressionner le destinataire et de l'impliquer dans son propos. Le phénomène d'identification que sous-tend la persuasion implique un discours fortement modalisé tant du point de vue de la modalité élocutive que de la modalité allocutive (cf. Charaudeau 1992). La description énonciative des prédicats décrit les modalités qu'impliquent leurs significations.

Nous avons catégorisé les prédicats du français d'un point de vue énonciatif. La catégorisation énonciative résulte des classes sémantiques des prédicats, car elles peuvent les impliquer dans un type de modalité. Nous distinguons trois sortes de description. La description interindividuelle concerne tous les prédicats en rapport avec une relation entre deux êtres humains. La description subjective concerne tous les prédicats en rapport avec l'intériorité d'un être humain. La description objective concerne tous les autres prédicats.

Les sous-catégories énonciatives regroupent des classes sémantiques. Par exemple, la sous-catégorie AFFECT subsume des classes comme JOIE (*être joyeux*, *être allègre*, *être euphorique*), PEUR (*être apeuré*, *être effrayé*, *être paniqué*), TRISTESSE (*être triste*, *être affligé*, *être éploré*), etc. Il s'ensuit qu'à partir de la classe sémantique d'un emploi prédicatif, on détermine sa catégorie énonciative en fonction de ses particularités syntaxiques. Par exemple, l'emploi adjectival *affligeant* a comme classe sémantique TRISTESSE, de telle

sorte que, d'un point de vue sémantico-énonciatif, il est sous-catégorisé en tant qu'AFFECT et, de ce fait, catégorisé en tant que DESCRIPTION SUBJECTIVE. L'adjectif peut être constitutif d'une simple assertion, c'est-à-dire une modalité délocutive, dans *On pleure quand cela est aussi affligeant* ou d'une modalité élocutive dans *C'est affligeant de faire cela* (Buvet 2011).

L'analyse et la description des prédicats adjectivaux, nominaux et verbaux est une contribution majeure au repérage des contenus propositionnels dans les textes et à leur interprétation automatique. L'étude de l'actualisation met en évidence le mode de fonctionnement des structures prédicat-argument dans les discours. Le concept d'emploi prédicatif synthétise les résultats des deux études en expliquant la nature des propriétés sémantiques d'un prédicat instancié dans un énoncé en fonction de ses propriétés morphosyntaxique et distributionnelles. La catégorisation énonciative des prédicats permet la transition entre sémantique lexicale et sémantique énonciative, afin d'aboutir à une interprétation globale des textes.

## 3. Ressources linguistiques

L'horizon applicatif de la théorie des trois fonctions primaires est la linguistique informatique. Cette dernière a pour finalité l'analyse et la compréhension automatique des textes numérisés et la mise en œuvre d'une linguistique outillée (Habert 2004). Sur le plan méthodologique, le développement d'applications permet de tester la validité des concepts utilisés dans le modèle de données. La linguistique informatique répond aux exigences suivantes: 1) observer les faits de langue étudiés pour en dégager des régularités généralisables; 2) exprimer ces régularités en éliminant le flou, l'implicite, le non-dit, les évidences allant de soi; 3) vérifier la cohérence de la formulation qui garantit l'objectivité et donc la reproductibilité de la démarche.

Le traitement de l'information, tel que nous le concevons, exploite des ressources linguistiques de qualité à large couverture. Il s'agit d'une approche dite linguistique. Elle est fondée sur le principe suivant: l'analyse sémantique des textes présuppose qu'ils soient enrichis d'informations métalinguistiques pertinentes. Ces dernières sont de deux ordres: morphosyntaxiques, d'une part, sémantiques, d'autre part. Trois sortes de ressources contribuent à l'enrichissement des textes: les dictionnaires électroniques, les grammaires locales et les corpus. Les deux premières le font directement, la dernière indirectement, puisque la fonction des corpus est de contribuer à l'élaboration et à la validation des dictionnaires et des grammaires locales.

L'approche linguistique est souvent complémentaire de l'approche mathématique. Cette dernière exploite essentiellement des outils statistiques et des méthodes probabilistes pour faire ressortir

des régularités lorsque les données à traiter sont quantitativement élevées. La complémentarité entre les deux approches, dite approche mixte, permet d'allier la finesse d'analyse de l'approche linguistique à la puissance des calculs de l'approche mathématique.

### 3.1. Dictionnaires électroniques

Les dictionnaires électroniques utilisés pour le traitement de l'information textuelle sont des dictionnaires morphosyntaxiques et des dictionnaires syntactico-sémantiques. Les dictionnaires du premier type sont dits MORFETIK; il en existe deux sortes: MORF-SIM (décrivant des unités monolexicales) et MORF-COM (décrivant des unités polylexicales) (cf. Mathieu-Colas *et al.* 2009, Mejri 1997). Les dictionnaires du second type sont dits DEESSE; il en existe quatre sortes: PRED-DIC (dictionnaires des emplois prédicatifs), ARGU-DIC (dictionnaires des arguments), ACTU-DIC (dictionnaires des actualisateurs) et ETHU-DIC (dictionnaire des êtres humains).

Le traitement de l'informatique textuelle implique une première étape consistant à identifier tous les mots d'un texte (phase de segmentation), puis à les normaliser (phase de lemmatisation) pour leur attribuer une catégorie grammaticale (phase de catégorisation). Pour ce faire, le système exploite des dictionnaires MORF-SIM et MORF-COM qui ont la forme de bases de données relationnelles comportant, dans différentes tables, la forme unique ou lemmatisée de toutes les unités lexicales du français ainsi que des règles de flexion qui expliquent leurs éventuelles variations (pluriels, conjugaisons, etc.). Les données ont été établies à partir de nombreuses sources lexicographiques. Un moteur de flexion produit l'ensemble de toutes les formes fléchies afin d'effectuer la lemmatisation et la catégorisation des mots d'un texte. Une base de règles est également exploitée afin de gérer les ambiguïtés catégorielles (par exemple, *porte* selon ses environnements s'analyse comme un substantif ou un verbe).

Parmi les différentes sortes de dictionnaires DEESSE, nous discutons uniquement de ceux du type PRED-DIC, car ce sont les seuls qui décrivent les prédicats d'affect. La nomenclature des dictionnaires PRED-DIC est constituée d'emplois prédicatifs. Les dictionnaires sont différenciés selon qu'il s'agit d'adjectifs prédicatifs (PRED-DIC-A), de noms (PRED-DIC-N) ou de verbes (PRED-DIC-V) (cf. Buvet 2009b).

La microstructure des dictionnaires est identique quelle que soit la forme des emplois décrits. Elle est constituée de descripteurs correspondant uniquement à des propriétés linguistiques. Ces descripteurs sont de deux sortes: les descripteurs de définition, d'une part, et les descripteurs de condition, d'autre part. Les descripteurs de définition correspondent aux propriétés sémantiques des emplois prédicatifs, ceux de condition à leurs propriétés formelles (cf. 1.1.).

Certains emplois prédicatifs correspondent à des prédicats qui sont instanciés sous une forme unique, d'autres le sont sous plusieurs formes. Dans le premier cas de figure, la forme de l'emploi est identique à celle de la racine prédicative, si ce n'est que cette dernière peut être accompagnée d'un indice numérique quand elle apparaît dans plus d'une structure prédicat-argument. C'est le cas de l'adjectif prédicatif *saumâtre* dans *Il est saumâtre*; ce dont rend compte le premier de ses descripteurs de définition:

| **Emploi** | **Racine prédicative** | **Classe sémantique** | **Type sémantique** | **Aspect inhérent** |
|---|---|---|---|---|
| *saumâtre* | saumâtre2 | DEPLAISIR | état | provisoire |

Dans le second cas de figure, soit les emplois prédicatifs sont sémantiquement équivalents, soit ils ne le sont pas totalement. Ainsi, les différentes instanciations de la structure a(i)m- (HUMAIN, HUMAIN) donnent lieu à des emplois prédicatifs dont les valeurs sont identiques: *Il est amoureux d'elle*; *Il éprouve de l'amour pour elle; Il l'aime*. Il s'ensuit qu'ils sont décrits avec les mêmes descripteurs de définition:

| **Emploi** | **Racine prédicative** | **Classe sémantique** | **Type sémantique** | **Aspect inhérent** |
|---|---|---|---|---|
| *amoureux* | a(i)m- | AMOUR | état | provisoire |
| *amour* | a(i)m- | AMOUR | état | provisoire |
| *aimer* | a(i)m- | AMOUR | état | provisoire |

Par contre, les différentes instanciations de la structure ha(ï|i)- (HUMAIN, HUMAIN) donnent lieu à des emplois prédicatifs dont les valeurs ne sont pas strictement identiques: *Il est haïssable*, *Il est haineux, Il éprouve de la haine pour elle*; *Il la hait*. Les descripteurs de définition font état de ces nuances:

| **Emploi** | **Racine prédicative** | **Classe sémantique** | **Type sémantique** | **Aspect inhérent** |
|---|---|---|---|---|
| *haineux* | ha(ï|i)- - | HAINE | état | permanent |
| *haine* | ha(ï|i)- - | HAINE | état | provisoire |
| *haïr* | ha(ï|i)- | HAINE | état | provisoire |

Le rôle des descripteurs de condition est de préciser quelles propriétés formelles sont nécessaires pour identifier un emploi prédicatif et lui attribuer automatiquement une étiquette sémantique. A titre d'exemple, nous indiquons quels sont les différents descripteurs de condition des emplois prédicatifs dans cette section:

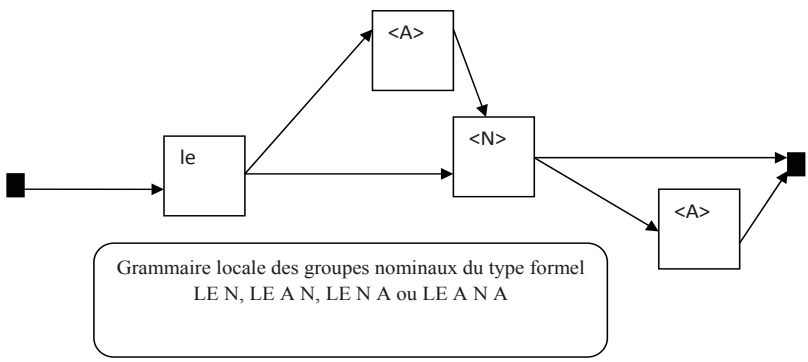
| Emploi | Construction | Distribution morpho-syntaxique | Distribution sémantique |
|---|---|---|---|
| *saumâtre* | X0 être A | X0 = GN | X0 = HUMAIN |
| *amoureux* | X0 être A (PREP1 X1) | X0 = GN<br>X1 = GN | X0 = HUMAIN<br>X1 = HUMAIN |
| *amour* | | X0 = GN<br>X1 = GN | X0 = HUMAIN<br>X1 = HUMAIN |
| *aimer* | | X0 = GN<br>X1 = GN | X0 = HUMAIN<br>X1 = HUMAIN |
| *haineux* | X0 être A | X 0= GN | |
| *haine* | | X0 = GN<br>X1 = GN | X0 = HUMAIN<br>X1 = HUMAIN |
| *haïr* | | X0 = GN<br>X1 = GN | X0 = HUMAIN<br>X1 = HUMAIN |

Dans la première colonne, X0 et X1 désignent les positons syntaxiques (sujet, premier complément) occupées par les arguments. Ces derniers sont décrits dans les deux autres colonnes.

### 3.2. Grammaires locales

Une grammaire locale décrit le cotexte d'une unité lexicale donnée en tant qu'ensemble de configurations de mots. Une grammaire locale est représentée par un graphe orienté. Elle est implémentée en tant qu'automate à états finis (le plus souvent, sous forme de transducteur, *i.e.* un automate qui reconnaît de l'information et qui en produit). Les informations métalinguistiques enregistrées dans les graphes au niveau des nœuds sont celles qui sont encodées dans les dictionnaires électroniques.

Une grammaire locale est représentée par un graphe comportant: 1) un nœud initial, un nœud final et un ensemble de nœuds intermédiaires; 2) des arcs qui relient les nœuds en fonction des configurations de mots de la grammaire locale (cf. Gross 1995).

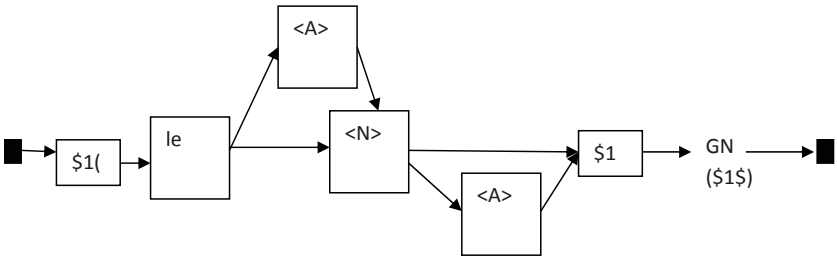


Les groupes nominaux *le chat, le beau chat, le chat gris* et *le beau chat gris* sont décrits par la grammaire locale ci-dessus.

Un automate à états finis est une expression régulière correspondant à une grammaire locale qui analyse des séquences de mots pour décider lesquelles sont conformes à l'une des configurations de la grammaire locale. Un transducteur à états finis est un automate à états finis qui associe une nouvelle information à l'information reconnue. Il permet également de remplacer les items d'une séquence reconnue par d'autres items. La première fonctionnalité est fondamentale pour l'étiquetage. La seconde fonctionnalité permet toutes sortes de manipulations des textes. Par exemple, on peut utiliser des transducteurs pour faire des représentations métalinguistiques, remplacer n'importe quelle phrase par sa construction canonique, etc.

Les automates et les transducteurs à états finis sont représentés sous forme de graphes afin de mieux les concevoir.

Qu'il s'agisse d'automates ou de transducteurs, les descripteurs métalinguistiques enregistrés dans les graphes au niveau des nœuds ou au niveau de l'information produite sont des patterns (des séquences de lettres) ou bien ceux qui sont encodés dans les dictionnaires électroniques. Autrement dit, un automate ou un transducteur exploite les dictionnaires électroniques en spécifiant les descripteurs de ces dictionnaires.




Représentation graphique d'un transducteur à états finis associant aux groupes nominaux du type formel LE N, LE A N, LE N A ou LE A N A l'étiquette GN

Les procédures d'étiquetage s'appuient sur la projection des dictionnaires électroniques sur les textes segmentés et catégorisés. Lorsqu'une forme est ambiguë, le système exploite des grammaires locales. L'étiquetage est une opération fondamentale pour l'interprétation automatique des textes

### 3.3. Les corpus

Les corpus exploités en Traitement Automatique des Langues sont généralement des textes bruts ou des textes annotés.

On peut les distinguer selon leur finalité: les corpus de travail, les corpus d'apprentissage, les corpus d'expérimentation, les corpus d'évaluation, etc.

L'élaboration d'outils performants pour effectuer les analyses linguistiques est fondée sur l'exploitation de trois types de corpus:

(i) le corpus d'investigation;

(ii) le corpus de test;

(iii) le corpus de validation.

Les trois sortes de corpus doivent porter sur des contenus différents pour éviter les solutions *ad hoc*. Le corpus d'investigation permet d'identifier les phénomènes linguistiques qui seront pris en compte et traités automatiquement. Compte tenu de ces phénomènes, il doit indiquer en premier lieu quelles ressources lexicographiques sont nécessaires au bon fonctionnement des outils d'analyse. Il doit permettre en second lieu de calibrer les transducteurs à états finis existants et, le cas échéant, de concevoir de nouveaux transducteurs.

La fonction du corpus de test est d'expérimenter les outils calibrés ou nouvellement développés afin de les corriger au fur et à mesure lorsqu'ils donnent lieu à du bruit (de l'information reconnue mais non pertinente) ou à du silence (de l'information pertinente mais non reconnue). Il doit permettre également de pointer d'éventuelles défaillances lexicographiques. L'exploitation du corpus de test doit accompagner au plus près la mise en place des outils afin d'anticiper les difficultés que peut entraîner leur utilisation après leur intégration dans une plateforme de travail.

Le corpus de validation fournit des résultats qui permettent de vérifier la qualité de l'ensemble des outils créés ou paramétrés. En cas d'invalidation, d'autres corpus d'investigation et de test sont utilisés pour améliorer les outils.

Les corpus de test et de validation sont des éléments essentiels à l'élaboration des dictionnaires et des grammaires locales. Ils doivent permettre d'établir puis de mesurer l'adéquation entre ces ressources et les phénomènes traités à différents niveaux d'analyse linguistique.

## 4. L'application TAEMA

L'importance de la part du vocabulaire dans l'apprentissage d'une langue et la possibilité d'améliorer les systèmes qui traitent l'information textuelle en implémentant des dictionnaires électroniques à large couverture constitue un point de convergence. Cela nous a conduits à développer un système d'aide à l'apprentissage qui met en

avant la dimension lexicale des langues. D'autant plus que le français langue seconde et la linguistique informatique ont des problématiques communes vis-à-vis des mots: le traitement de la polysémie, de la polymorphie et du figement sont les mêmes pour un apprenant ou un système opérant sur des données linguistiques. La nécessité de prendre en compte conjointement les propriétés morphologiques, syntaxiques et sémantiques des mots pour expliquer leur mode de fonctionnement est valable dans l'un et l'autre cas.

Le système d'aide à la rédaction TAEMA produit des phrases centrées sur le vocabulaire affectif du français. La particularité du dictionnaire PRED-DIC permet au système non seulement d'indiquer tout le vocabulaire en rapport avec un type d'affect donné (par exemple une émotion ou un sentiment particulier) mais aussi toutes les constructions associées avec ce vocabulaire. Au final, après avoir fait une requête relative à un type d'affect, l'utilisateur a la possibilité de choisir parmi un ensemble de phrases équivalentes à celle qui lui semble la plus adéquate pour sa production écrite.

Le prototype développé est susceptible de s'adresser aux apprenants du français langue seconde. Il convient de préciser que l'utilisation de l'interface nécessite une connaissance sommaire du français. L'usage du logiciel présuppose que les utilisateurs sachent exprimer au moins la phrase prototypique associée à la notion et l'environnement spécifié (par exemple *Luc aime Léa* pour l'affect AMOUR). La finalité du logiciel est de lui proposer toutes les paraphrases possibles (*Luc a le béguin pour Léa, Luc en pince pour Léa, Luc est fou de Léa*, etc.)

Du point de vue de l'utilisateur du système, il s'agit d'indiquer un type d'affect donné et des éléments contextuels afin d'obtenir toutes les phrases canoniques du français qui sont en rapport avec les indications fournies. Le système permet à un utilisateur de sélectionner (sous forme de menus déroulants ou en les spécifiant en langue naturelle) les différents éléments d'une phrase à produire. Nous détaillons les différents écrans qui permettent à l'utilisateur de formuler sa requête.

Le premier écran présente: (i) la finalité du projet; (ii) un mode d'emploi; (iii) un bouton 'entrée' qui permet d'accéder au système.

Le deuxième écran invite l'utilisateur à choisir un type d'affect dans un champ de saisie à l'aide d'un menu déroulant. Il s'agit de l'appellation de l'une des sous-classes de prédicats d'affect qui figurent dans le dictionnaire électronique utilisé par le système. Cette page comporte également trois autres champs de saisie qui ne sont pas activés à ce niveau d'utilisation. Le premier concerne la personne qui ressent l'affect. Le second a trait à l'éventuelle personne, ou entité, impliquée dans l'affect. Le troisième permet de conjuguer le prédicat à un temps grammatical simple ou complexe.

Le troisième écran rend actifs en partie ou en totalité les champs de saisie autres que celui déjà rempli. Selon que le prédicat est monadique ou dyadique, le nombre de champs accessibles varie. Des options permettent d'en stipuler les contenus soit par des menus déroulants soit en en spécifiant directement les informations attendues.

Le quatrième écran donne tous les résultats relatifs à la requête de l'utilisateur de telle sorte qu'il puisse choisir une des phrases relatives à l'affect spécifié et les éléments contextuels qu'il a précisés. Le prototype propose 60 concepts relatifs aux affects et peut générer plus de 3000 phrases simples. Par exemple, comme paraphrase de *J'ai peur*¸ TAEMA propose: *Je suis apeuré*, *Cela m'apeure*, *Je balise, Je suis effaré, Cela m'effare, Je suis effarouché, Cela m'effarouche*, *Je suis effrayé*, *Cela m'effraie*, *Je ressens de l'effroi*, *Je ressens de épouvante, Je suis épouvanté, Cela m'épouvante*, *Je suis horrifié, Cela m'horrifie, Quelle frayeur!, J'ai la frousse, Je mouille, Je panique, Cela me panique, Je suis paniqué, J'ai la pétoche, Je suis peureux, Je ressens de la terreur, Je suis terrifié, Cela me terrifie, Je suis terrorisé, Cela me terrorise, J'ai le trac, J'ai la trouille*.

## Références bibliographiques

André, C. (2009), *Les états d'âme. Un apprentissage de la sérénité*, Odile Jacob, Paris.

Anscombre, J.-C. (1995), « Morphologie et représentation événementielle: le cas des noms de sentiment et d'attitude », *Langue française* 105, p. 40-54.

Anscombre, J.-C. (1996), « Noms de sentiment, noms d'attitude et noms abstraits », *in* Flaux, N., Glatigny, M. et Samain, D. (dir.), *Les noms abstraits, histoire et théories*, Presses Universitaires du Septentrion, Lille, p. 257-273.

Benveniste, E. (1966), *Problèmes de linguistique générale*, Gallimard, Paris.

Buvet, P.-A. (2009a), « Des mots aux emplois: la représentation lexicographique des prédicats », *Le Français Moderne* 77/1, p. 83-96.

Buvet, P.-A. (2009b), « Quelles procédures d'étiquetage pour la gestion de l'information textuelle électronique? », *L'information grammaticale* 122, p. 40-48.

Buvet, P.-A. (2011), « Catégorisation sémantico-énonciative du lexique à partir d'un dictionnaire électronique », in *Os di.ci.o.nă.rios Fontes, métotodos e novas technologias*, Instituto de Letras da Universidade Federal da Bahia, p. 75-96.

Buvet, P.-A. (2012), « Traitement automatique du discours rapporté », *Actes du colloque JADT 2012*, Université de Liège.

Buvet, P.-A. (à paraître), « Des unités lexicales aux unités discursives: la catégorisation sémantico-énonciative des prédicats », in *L'unité en sciences du langage*, *Actualité scientifique*, AUF.

Buvet, P.-A., Girardin, C., Gross, G. et Groud, C. (2005), « Les prédicats d'affect », *LIDIL* 32, p. 125-143.

Buvet, P.-A. et Grezka, A. (2007), « Élaboration d'outils méthodologiques

pour décrire les prédicats du français », *Lingvisticae Investigationes* 30/2, p. 217-245.
Buvet, P.-A. et Issac, F. (2006), « TAEMA: Traitement Automatique de l'Ecriture de Mots Affectifs », *Verbum ex machina*, 2, Presses Universitaires de Louvain, Louvain-la-Neuve, p. 856-867.
Charaudeau, P. (1992), *Grammaire du sens et de l'expression*, Hachette, Paris.
Cresseils, D. (2000), « L'emploi résultatif de être + participe passé en français », *Cahiers Chronos* 6, p.133-142.
Cyrulnik, B. (1991), *De la parole comme une molécule*, Éditions ESHEL.
Gross, M. (1995), « Une grammaire locale de l'expression des sentiments », *Langue française* 105, p. 70-87.
Habert, B. (2004), *Instruments et ressources électroniques pour le français*, Ophrys, Gap-Paris.
Harris, Z. S. (1976), *Notes du cours de syntaxe*, Seuil, Paris.
James, W. (1894), "Discussion: The physical basis of emotion", *Psychological Review* 1/5, p.516-529.
Mathieu-Colas, M., Buvet, P.-A., Cartier, E., Issac, F., et Mejri, S. (2009), « Morfetik, ressource lexicale pour le TAL », *Actes de TALN 2009*, Senlis (http://www-lipn.univ-paris13.fr/taln09/pdf/TALN_26.pdf).
Mattelart, A. et Mattelart, M. (2004), *Histoire des théories de la communication*, Éditions La Découverte, Paris.
Mejri, S. (1997), *Le figement lexical: descriptions linguistiques et structuration sémantique*, Publications de la Faculté des lettres de la Manouba, Tunis.
Mejri, S. (2009), « Le mot: problématique théorique », *Le Français Moderne* 77/1, p. 68-82.
Melis, L. (2003), *La préposition en français*, Ophrys, Paris-Gap.
Nazarenko, A. (2000), *La cause et son expression en français*, Ophrys, Paris.
Reboul, A. et Moeschler, J. (1998), *La pragmatique aujourd'hui*, Seuil, Paris.
Rosier, L. (2008), *Le discours rapporté en français*, Ophrys, Paris.
Shannon, C. (1948), "A Mathematical Theory of Communication", *Bell System Technical Journal* 27, p. 379-423.
Tutin, A., Novakova, I., Grossmann, F. et Cavalla, C. (2006), « Esquisse de typologie des noms d'affect à partir de leurs propriétés combinatoires », *Langue française* 150, p. 32-49.
Van de Velde, D. (1995), *Le spectre nominal*, Peeters, Louvain-Paris.