

---

## Traitement des lexies d'émotion dans les corpus et les applications d'*EmoBase*

Sascha Diwersy, Vannina Goossens, Anke Grutschus, Beate Kern, Olivier  
Kraif, Elena Melnikova et Iva Novakova

---



### Édition électronique

URL : <http://journals.openedition.org/corpus/2537>

DOI : [10.4000/corpus.2537](https://doi.org/10.4000/corpus.2537)

ISSN : 1765-3126

### Éditeur

Bases ; corpus et langage - UMR 6039

### Édition imprimée

Date de publication : 1 décembre 2014

Pagination : 269-293

ISSN : 1638-9808

### Référence électronique

Sascha Diwersy, Vannina Goossens, Anke Grutschus, Beate Kern, Olivier Kraif, Elena Melnikova et Iva Novakova, « Traitement des lexies d'émotion dans les corpus et les applications d'*EmoBase* », *Corpus* [En ligne], 13 | 2014, mis en ligne le 01 mai 2015, consulté le 08 septembre 2020. URL : <http://journals.openedition.org/corpus/2537> ; DOI : <https://doi.org/10.4000/corpus.2537>

---

## Traitement des lexies d'émotion dans les corpus et les applications d'*EmoBase*

Sascha DIWERSY<sup>(1)</sup>, Vannina GOOSSENS<sup>(2)</sup>,  
Anke GRUTSCHUS<sup>(1)</sup>, Beate KERN<sup>(3)</sup>, Olivier KRAIF<sup>(4)</sup>,  
Elena MELNIKOVA<sup>(4)</sup>, Iva NOVAKOVA<sup>(4)</sup>

(1) Université de Cologne,

(2) ICAR/ENS de Lyon,

(3) Université de Rostock,

(4) LIDILEM/Université Grenoble-Alpes

### 1. Fondements théoriques pour la construction d'une ressource et de ses outils d'exploitation

La plate-forme *EmoBase* (<http://emolex.u-grenoble3.fr/emoBase>) a été créée dans le cadre du projet franco-allemand ANR-DFG EMOLEX (2009-2013) dont le principal objectif a été d'étudier le lexique des émotions dans cinq langues européennes (français, espagnol, allemand, anglais, russe)<sup>1</sup>. Cette plate-forme rassemble des corpus (cf. section 2), des applications destinées à leur interrogation (*EmoConc*, cf. sections 4.1 à 4.3) ainsi qu'une interface permettant d'accéder aux résultats des analyses du lexique des émotions (*EmoLing*, cf. section 4.4)<sup>2</sup>.

La méthodologie mise en place pour l'analyse du lexique des émotions s'appuie sur deux conceptions théoriques, à savoir l'étude du « sens≈concept » (approches « représentationnistes ») et l'étude du « sens≈usage » (approches « instrumentalistes »)<sup>3</sup>. La première conception, déductive et taxinomique, a pour objectif l'étude des combinaisons phraséologiques (ou collocations) que Hausmann & Blumenthal (2006) définissent comme

---

<sup>1</sup> Le projet ([www.emolex.eu](http://www.emolex.eu)) a réuni des équipes de chercheurs des universités Stendhal (Grenoble 3), de Cologne et d'Osnabrück.

<sup>2</sup> Outre les applications mentionnées, *EmoBase* contient également une application didactique (*EmoProf*), dont la présentation dépasserait le cadre de cet article.

<sup>3</sup> Pour plus de détails à ce sujet, cf. Keller (1995).

étant des associations, hors contexte, entre une base (**pivot**) comme *examen, célibataire, colère*) et un **collocatif** (*passer, endurci, bouffée de*). Ces combinaisons correspondent aux expressions semi-figées *donner le sein, donner un cours* qui se distinguent des constructions libres (*donner un livre*) ou figées (*donner du fil à retordre*) (Mel'čuk et al. 1995). La deuxième conception, inductive et systémique, qui relève du contextualisme britannique (Firth 1957, Sinclair 1991, Hoey 2005), s'intéresse aux combinaisons récurrentes entre les mots en contexte. Il s'agit d'une approche textuelle, différente d'une approche lexicographique, s'appuyant sur l'usage à travers de grands corpus informatisés.

A partir de ces deux orientations théoriques, une méthodologie spécifique à l'étude du lexique des émotions a été élaborée, combinant deux approches jusqu'à présent exploitées séparément. D'une part, l'analyse du lexique au moyen de méthodes lexico-statistiques (Diwersy & Kraif 2013) permet d'établir les accompagnateurs spécifiques ou préférentiels (co-occurents) des lexies d'émotion sur grands corpus. D'autre part, grâce à l'étude systématique de la combinatoire (Tutin et al. 2006, Novakova & Tutin 2009) sont identifiées les propriétés syntaxico-sémantiques des associations lexicales (*tomber amoureux, to find happiness* 'trouver le bonheur', *gluboko uvažat* 'respecter profondément').

## 2. Les corpus comparables et parallèles multilingues intégrés dans *EmoBase*

*EmoBase* rassemble des corpus dans cinq langues : le français, l'allemand, l'anglais, l'espagnol et le russe. Elle est composée de deux types de corpus : comparables et parallèles alignés. Les corpus comparables comprennent environ 140 M de mots par langue : des textes journalistiques<sup>4</sup> pour un total d'environ

---

<sup>4</sup> Il s'agit d'articles de presse quotidienne nationale et régionale parus dans les années 2000 : par exemple, pour le corpus français, les articles sont issus de *Libération, Le Monde, Le Figaro* et *Ouest-France* pour les années 2007 et 2008. Les journaux de même type dans les autres langues datent des mêmes années.

120 M de mots, et des textes littéraires représentant 20 M de mots (pour l'essentiel des romans des années 1950 à nos jours).

Le corpus parallèle a une taille d'environ 78 M de mots au total et comprend uniquement des textes littéraires (des romans du XIX<sup>e</sup> et du XX<sup>e</sup> siècle, la plus grande part étant constituée de romans contemporains) qui ont été alignés avec leur traduction respective à l'aide du programme *Alinea*.

Les corpus ont été étiquetés et annotés syntaxiquement et sont interrogeables depuis l'interface *EmoConc*<sup>5</sup>, qui sera décrite dans la section 4. Cette interface permet deux modes d'interrogation et d'affichage des données du corpus.

Le premier mode d'interrogation permet d'afficher des concordances comportant toutes les occurrences d'une entité recherchée dans le corpus (un lemme, une forme, des pivots complexes, etc.). Le deuxième mode d'interrogation est plus complexe. Il est fondé sur le calcul probabiliste *log likelihood*<sup>6</sup>, permettant de filtrer et de pondérer les cooccurrents d'un mot ou d'une forme donnés.

Les données extraites ont été ensuite codées selon une grille sémantique et syntaxique appliquée aux cinq langues du projet. Les résultats de cette analyse linguistique sont accessibles via une deuxième interface, nommée *EmoLing*. (cf. 4, 5.1 et 5.2).

Les utilisateurs d'*EmoBase* ont donc accès à la fois aux données brutes (les textes des corpus, sous forme de concordances), à des informations statistiques concernant la cooccurrence, et à une description linguistique détaillée (grille sémantique, structures actanciennes) ciblée sur le lexique des émotions. Grâce à une interface d'interrogation simple et ergonomique, les utilisateurs peuvent extraire tout type de lexique (mots, collocations) sous forme de concordances. A partir des concordances, on peut avoir accès à la phrase dans laquelle se trouvent ces occurrences, ainsi qu'à un contexte légèrement élargi (de deux à trois phrases avant ou après la phrase extraite), ainsi qu'aux métadonnées (source, date, auteur). Afin de permettre à

---

<sup>5</sup> L'interface a été créée par Olivier Kraif, Sascha Diwersy et Sylvain Hatier.

<sup>6</sup> Pour une présentation des différents calculs statistiques, cf. Blumenthal *et al.* 2005.

l'utilisateur de mener à bien ses propres analyses en partant de la description linguistique proposée sur *EmoLing*, nous souhaitons, dans un premier temps, donner un aperçu de la méthodologie de sélection des lexies. Dans un deuxième temps (cf. 5), nous présenterons les différents niveaux d'analyse et proposerons des exemples d'exploitation.

### **3. Méthodologie de la sélection des champs et des lexies d'émotion**

L'analyse intra- et interlinguistique des lexies d'affect a été fondée sur les corpus Emolex décrits dans la précédente section. Le domaine des affects étant très large, il a été procédé à une sélection de champs sémantiques des émotions et aussi des lexies qui les composent selon différents critères quantitatifs et qualitatifs.

#### **3.1 Sélection des champs analysés**

Neuf champs sémantiques d'émotion ont été retenus, incluant des affects de polarité positive (*respect, joie, admiration*), négative (*colère, déception, jalousie, mépris, tristesse*) et neutre (*surprise*) (cf. Grutschus *et al.* 2013). Nous avons également pris en compte différents types d'affects : *réactifs* impliquant prototypiquement un expérienceur plutôt qu'un agent (*déception, surprise, colère, joie*), *interpersonnels* orientés vers un objet humain (*respect, mépris, admiration, jalousie*) et enfin des affects renvoyant à des états comme *tristesse* (cf. Tutin *et al.* 2006). La sélection comprend aussi bien des champs traités plus fréquemment dans les analyses linguistiques (p. ex. *colère* ou *tristesse*) que des champs moins souvent étudiés (p. ex. *surprise* ou *admiration*).

Plusieurs types d'analyses linguistiques ont été réalisés au cours du projet. Les analyses sémantiques et syntaxiques ont été faites sur la totalité des champs. Les analyses actanciennes et discursives<sup>7</sup> ont été effectuées seulement pour une partie des champs (figure 1) :

---

<sup>7</sup> L'analyse discursive n'a pas été formalisée et n'apparaît donc pas dans les bases de données interrogeables en ligne. Les résultats sont présentés dans plusieurs publications, p. ex. Novakova & Sorba (2013) et Grutschus & Kern (à par.).

### Traitement des lexies d'émotion

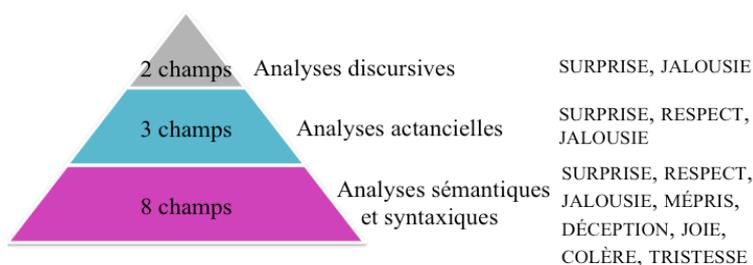


Figure 1 : les niveaux d'analyse et les champs traités

### 3.2 Sélection des lexies analysées

Une fois les champs sémantiques sélectionnés, il a été procédé à la sélection des lexies constitutives de chaque champ pour les cinq langues. Cette étape a été très importante car elle devait garantir la comparabilité interlinguistique. Il s'agissait aussi de prendre en compte les fréquences d'emploi des différentes lexies, conformément aux méthodes statistiques choisies. A ce stade, il était, enfin, nécessaire d'assurer une couverture assez large et représentative des lexies pour chaque champ. La sélection des lexies par champ s'est faite en deux phases.

#### 3.2.1 Identification des lexies « candidates » pour chaque champ

Le français étant la langue pivot, ont été d'abord identifiées les lexies « candidates » pour chaque champ (par ex. *colère*, *indignation*, *exaspération* pour le champ COLÈRE). Les équivalents allemands, espagnols, anglais et russes ont été déterminés suite à la consultation de différentes ressources lexicographiques<sup>8</sup> ainsi qu'à des sondages auprès de locuteurs natifs. En partant de

<sup>8</sup> Les dictionnaires suivants ont été utilisés :

Français: *Trésor de la langue française informatisé*, *Petit Robert*  
Allemand : *Duden*, <http://wortschatz.uni-leipzig.de>, <http://dwds.de>  
Espagnol : *Diccionario de la Real Academia Española*, *Diccionario de uso del español*  
Anglais : *The Oxford Dictionary of English*, *Oxford Thesaurus of English*, *Collins English Thesaurus*, *Webster's new dictionary of synonyms*  
Russe : *Словарь русского языка (под ред. А. П. Евгеньевой, febweb.ru)*, *Словарь русского языка (под ред. Н. Ю. Шведовой)*, *Толковый словарь русского языка (под ред. Д. Н. Ушакова)*, *Словарь русских синонимов (под ред. Н. Абрамова)*.

la lexie de base française et de ses équivalents dans les autres langues, des synonymes nominaux, verbaux et adjectivaux ont été relevés dans chaque langue.

Une attention particulière a été portée à la forte polysémie des unités lexicales abstraites. Les lexies candidates doivent obligatoirement avoir une acception ‘affect’ (qui est parfois proche d’acceptions de type ‘attitude’ ou ‘qualité’) que nous avons identifiée à l’aide de plusieurs critères et tests (applicables essentiellement aux substantifs). Pour obtenir le statut de candidat, il a été jugé suffisant qu’une lexie remplisse au moins un des critères suivants (cf. Goossens 2011) :

- 1 Les définitions lexicographiques de la lexie en question la décrivent comme « sentiment », « émotion », « affect », « état affectif », etc.
- 2 La lexie apparaît en cooccurrence avec *ressentir*, *éprouver*, *sentiment (de)*, etc.
- 3 La lexie accepte la cooccurrence avec un actant exprimant une cause ou un objet, ce qui la différencie des noms de qualité.
- 4 La lexie répond positivement à l’un des tests d’intériorité. Un premier test stipule que les affects, qui renvoient à un ressenti intérieur, ne sont pas compatibles avec *je trouve que* contrairement aux attitudes qui donnent lieu à un ressenti extérieur (*\*je trouve Max amoureux*) (cf. Ducrot 1975, Anscombe 1995, Buvet *et al.* 2005). Un deuxième test se fonde sur des constructions locatives. Celles-ci peuvent être dynamiques (*Il fut envahi par un désespoir profond*), ce qui montre que l’affect vient habiter l’expérienceur depuis l’extérieur. Elles peuvent être aussi statiques en plaçant l’expérienceur à l’intérieur de l’affect (*être en colère*, *être dans la peine*). Cette dernière construction s’oppose à la construction spécifique des qualités qui place la propriété à l’intérieur du sujet (*Il y a chez cet homme / en lui une grande générosité*) (cf. Flaux & Van de Velde 2000).
- 5 La lexie est caractérisée par un rapport spécifique au temps : les affects, qui sont transitoires, acceptent la combinaison

avec l'adverbe *constamment* (*Il est constamment apeuré*) (Buvet *et al.* 2005), contrairement aux qualités<sup>9</sup>.

- 6 La lexie se combine avec certains collocatifs spécifiques à l'interprétation d'affect, tels que des collocatifs aspectuels (surtout ponctuels : *élan, transport, bouffée, flambée*), des collocatifs qui marquent l'envahissement par l'affect (*envahir, submerger*) et des collocatifs qui expriment des manifestations physiques (*trembler*).

### 3.2.2 Evaluation et validation du statut de « membre du champ »

Toutes les lexies candidates ainsi obtenues ont été soumises à des vérifications de fréquence pour assurer une fréquence suffisante permettant l'application de certains calculs statistiques.

Les lexies qui, dans le corpus, ont une fréquence absolue inférieure à 100 ont ainsi été exclues. Pour les unités lexicales polysémiques, il a fallu procéder à une désambiguïsation à l'aide du concordancier (*EmoConc*). Pour des raisons de faisabilité, nous avons vérifié 100 occurrences de la lexie candidate afin d'identifier le nombre d'occurrences avec l'acception « affect ». Celui-ci a permis d'extrapoler le nombre des occurrences avec l'acception « affect » sur la totalité des occurrences. Si, après extrapolation, le nombre des occurrences avec l'acception « affect » dépasse la barre des 100, la lexie peut intégrer le champ étudié.

## 4. Traitement et interrogation des corpus

### 4.1 Le concept de lexicogramme

Pour étudier le profil combinatoire des lexies retenues selon le procédé décrit, nous avons développé une application nommée *EmoConc*, accessible en ligne à partir d'*EmoBase*. Celle-ci est fondée sur le concept de *lexicogramme* utilisé dans le logiciel WebLex (Heiden & Tournier 2000) : il s'agit d'établir, pour un pivot donné, la liste de ses cooccurrents les plus fréquents, à gauche et à droite, en intégrant à la fois les fréquences de co-occurrence et des mesures d'association statistiques (telles que le rapport de vraisemblance). Dans *EmoConc*, l'utilisateur peut

---

<sup>9</sup> Ce test n'est pas suffisant pour distinguer affect et attitude, les attitudes étant également transitoires.

définir lui-même les unités visées dans la cooccurrence : formes, lemmes, catégories morphosyntaxiques, traits additionnels (p. ex. sémantiques), relations syntaxiques (dans le cas des *colligations*) ou toute combinaison de ces informations. Il est également possible de préciser des contraintes sur le contexte d'un pivot.

Par exemple, on peut étudier la cooccurrence de noms déterminés par un adjectif possessif, ce qui s'avère un critère extrêmement filtrant pour caractériser des emplois dans le champ des noms d'émotion (*son trouble évident vs ces troubles évidents*). On parle dans ce cas de *pivots complexes*. Cette flexibilité dans la définition des unités nous semble importante pour permettre à l'utilisateur d'ajuster la focale de ses observations en allant du général au particulier (ou vice-versa), de préciser des contraintes pour désambiguïser certains contextes, et de combiner les aspects lexicaux et syntaxiques dans ses observations. En ce qui concerne *l'espace de cooccurrence*, qui conditionne les points de rencontre entre pivot et collocatifs et la manière de les dénombrer, nous avons opté, à l'instar de Kilgariff & Tugwell (2001) ou de Charest *et al.* (2010), pour la *cooccurrence syntaxique*, qui repose sur les relations fonctionnelles du type sujet, complément d'objet, modifieur, etc. Evert (2009 : 1223) signale l'intérêt de ce type de cooccurrence en terme de bruit et de silence, par rapport aux simples cooccurrences de surface : « [...] unlike surface cooccurrence, it does not set an arbitrary distance limit, but at the same time introduces less 'noise' than textual cooccurrence ».

Pour la cooccurrence syntaxique, nous exploitons les relations de dépendances obtenues grâce à différents analyseurs : XIP pour l'anglais (Aït-Mokhtar *et al.* 2002), Connexor pour l'allemand, le français et l'espagnol (Tapanainen & Järvinen 1997), DeSR pour le russe (Attardi *et al.* 2007), basé sur un modèle stochastique créé à partir du corpus arboré SyntagRus (Nivre *et al.* 2008). Nous avons par la suite complété ces relations pour obtenir des dépendances plus pertinentes sur le plan sémantique (p. ex. sujet profond dans les constructions passives, etc.).

Enfin, l'espace de cooccurrence peut porter sur tous les types de relations de dépendances, ou seulement certaines rela-

tions précisées par l'utilisateur. Avec le modèle de cooccurrence ainsi défini, on peut viser des aspects très génériques de la combinatoire (par exemple : quels sont les principaux collocatifs de la forme *surprise* toutes relations confondues) ou beaucoup plus spécifiques et circonscrits (par exemple : quels sont les principaux collocatifs verbaux à l'imparfait du nom lemmatisé *surprise* en tant qu'objet direct). Le tableau 1 montre un tel lexicogramme.

Tableau 1. Extrait du lexicogramme pour le nom lemmatisé *surprise* pris en tant qu'objet direct (*f*=fréquence de cooccurrence, *f1*=fréquence de *l1*, *f2*=fréquence de *l2*)

<b>l1</b>	<b>l2</b>	<b>f</b>	<b>f1</b>	<b>f2</b>	<b>loglike</b>
surprise_N	créer_V	614	2098	21658	4548,4333
surprise_N	réserver_V	230	2098	2869	2143,50164
surprise_N	avoir_V	484	2098	423602	627,503103
surprise_N	constituer_V	94	2098	13778	406,792757
surprise_N	éviter_V	43	2098	16296	109,29478
surprise_N	manifester_V	22	2098	2424	106,621896

#### 4.2 Concordances et représentations graphiques

A partir de ces lexicogrammes, *EmoConc* offre différentes modalités d'exploration.

- Pour l'analyse linguistique, le « retour au texte » est indispensable : un simple clic sur un collocatif permet de retrouver, sous forme de concordance, tous les contextes de cooccurrence avec le pivot. Par ailleurs l'accès aux concordances peut être effectué directement, sans passer par un lexicogramme, par l'écriture d'une requête simple ou complexe dans un formalisme proche de celui de CQP, décrit dans Kraif (2008).
- Pour comparer de manière synthétique divers profils combinatoires, nous proposons d'identifier les lexicogrammes à des points dans un espace vectoriel, en ne retenant que la mesure jugée la plus pertinente (fréquence, loglike, t-score,

etc.). Il est dès lors possible d'utiliser des méthodes d'analyse de données pour visualiser les similarités entre pivots : analyse factorielle des correspondances (AFC), échelonnement multidimensionnel (MDS) ou classification hiérarchique ascendante (hClust), cf. figures 2a & b.

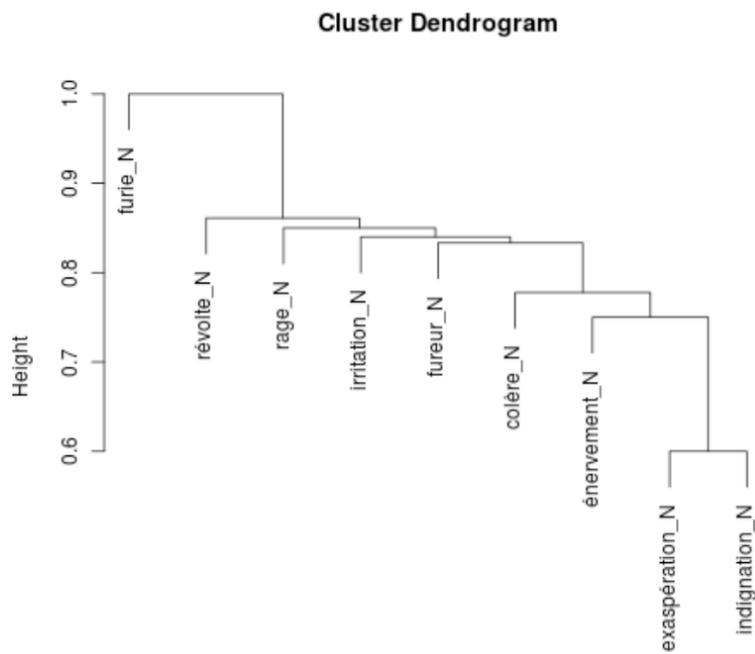


Figure 2a. Classification hiérarchique (domaine sémantique de la COLÈRE)

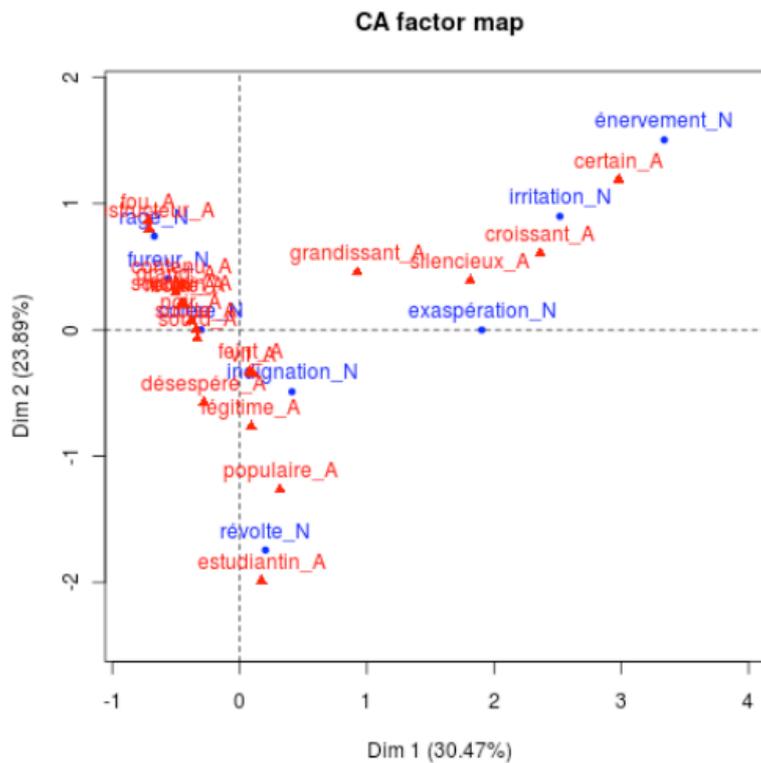


Figure 2b. AFC (domaine sémantique de la COLÈRE)

#### 4.3 Extraction d'unités polylexicales

La possibilité de calculer les lexicogrammes d'un pivot complexe, représentant un sous-arbre syntaxique, nous permet d'étendre ce sous-arbre en y adjoignant ses collocatifs les plus significatifs. En appliquant cette méthode itérativement, *EmoConc* peut extraire à la demande, sans précalcul, des expressions récurrentes de longueur  $n$  autour d'un pivot donné. Ces expressions récurrentes dépassent le simple cadre des segments répétés ou paquets lexicaux : elles représentent de véritables sous-arbres syntaxiques récurrents, susceptibles de se réaliser, de différentes manières, en surface dans les textes. L'exemple de la

figure 3 montre un tel sous-arbre, extrait automatiquement en partant du pivot admiration :

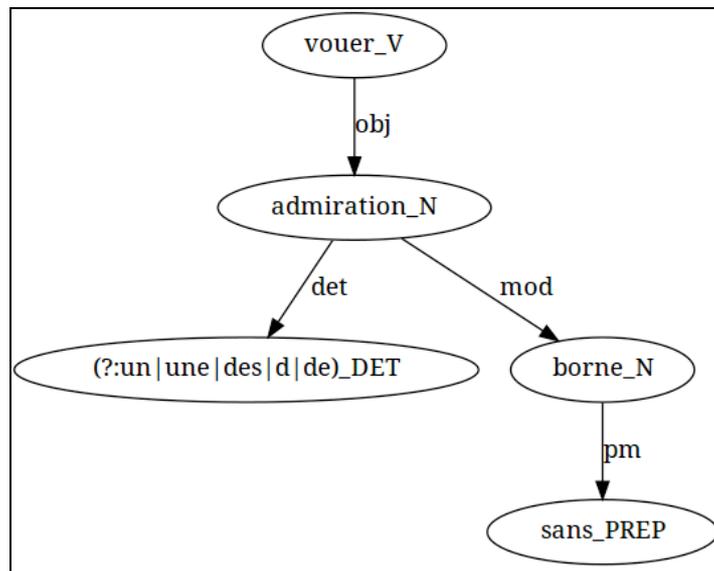


Figure 3. Extraction automatique du sous-arbre correspondant à « vouer une admiration sans bornes »

#### 4.4 EmoLing : EmoGrammes et structures actancielles

L'application *EmoLing* est destinée à rendre accessible la base de données lexicales comprenant les résultats des analyses linguistiques qui seront décrites ci-dessous (cf. 5.). L'interface est composée de deux formulaires, l'un pour les requêtes portant sur les structures actanciennes (cf. figure 4), l'autre pour la sélection et la visualisation des cooccurrences lexico-syntaxiques des lexèmes étudiés, les *EmoGrammes* (cf. figure 5).

## Traitement des lexies d'émotion

The screenshot shows a web interface titled 'Emogrammes Structures actancielles'. Under the heading 'Sélection', there are several input fields: 'Langues' with 'Français' selected, 'Champs sémantiques' with 'surprise' selected, 'Pivots - Lemmes', 'Pivots - Catégories', 'Emplois', 'Rôles sémantiques', 'Fonctions syntaxiques', 'Type d'actant', and 'Syntagmes'. Each field has a 'Select Some Options' button. A red 'Afficher' button is located at the bottom left.

Figure 4. Formulaire pour les requêtes portant sur les structures actancielles

The screenshot shows a web interface titled 'Emogrammes Structures actancielles'. Under the heading 'Sélection des entrées', there are several input fields: 'Langues' with 'Français' selected, 'Champs sémantiques' with 'SURPRISE' selected, 'Pivots - Lemmes', 'Pivots - Catégories' with 'N' selected, 'Collocatifs - Lemmes', 'Collocatifs - Catégories', 'Relations syntaxiques', 'Dimensions sémantiques', and 'Valeurs de dimension'. Each field has a 'Select Some Options' button. Below this, under 'Critères de groupement', there are 'Éléments' (Langues, Champs, Lemmes - Pivots) and 'Variables' (Dimensions) dropdowns. A checkbox 'Dissocier les dimensions sémantiques combinées' is present. A red 'Afficher' button is at the bottom left.

Figure 5. Formulaire pour la sélection et la visualisation des cooccurrences lexico-syntaxiques (application EmoGrammes)

Cette visualisation des données cooccurentielles s'effectue sous forme d'un tableau et de plusieurs types de graphiques (graphiques à secteurs, diagrammes de Pareto, treemaps). Elle peut être configurée de façon à regrouper les cooccurrences sélectionnées selon différents paramètres (langue, classe sémantique, lexème) et à proportionner, par rapport aux items issus de ce regroupement, le poids collocationnel<sup>10</sup> revenant à des catégories définies selon la dimension sémantique instanciée par les collocations concernées ou encore les constructions syntaxiques impliquées. En guise d'exemple, les diagrammes de Pareto donnés par les figures 6 et 7 représentent le poids respectif des dimensions sémantiques réalisées par les collocatifs des substantifs

<sup>10</sup> Le poids collocationnel se calcule sur la base des scores *log-likelihood* normalisés par rapport à leurs maxima et minima pour une entité donnée. Sur la procédure de calcul correspondante cf. Diwersy (2012 : 76 sq).

français *surprise* et *étonnement*. Ils montrent, entre autres, pour le profil combinatoire de *surprise*, la prédominance de la dimension « polarité », absente du profil d'*étonnement*.

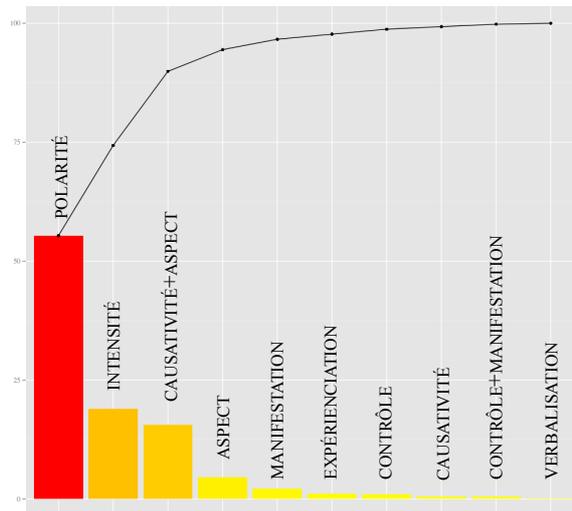


Figure 6. Poids collocationnel des dimensions sémantiques dans le profil combinatoire du nom *surprise*

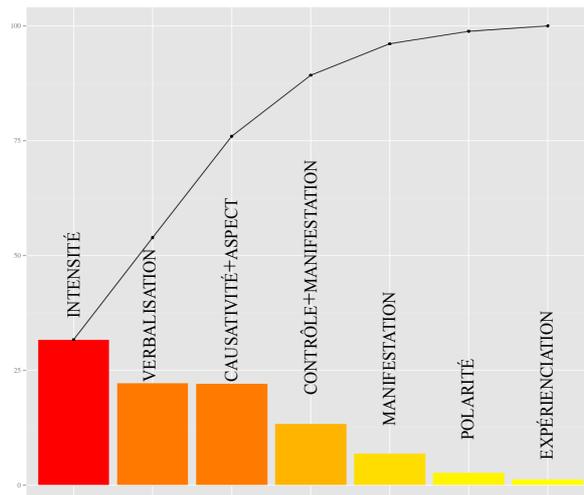


Figure 7. Poids collocationnel des dimensions sémantiques dans le profil combinatoire du nom *étonnement*

## 5. Traitement et analyse des données

Dans ce qui suit, nous allons d'abord illustrer sur quels niveaux et selon quels critères les lexies d'affect contenues dans l'application EmoLing ont été analysées. Sur chaque niveau d'analyse, un exemple d'exploitation d'EmoLing sera proposé.

### 5.1 Grille sémantique

Une partie des cooccurrents spécifiques<sup>11</sup> des lexies d'affect sélectionnées et qui ont été recensés dans les lexicogrammes générés dans *EmoConc* (cf. 4.1), ont été codés à l'aide d'une grille composée de huit dimensions sémantiques spécifiques du lexique des affects (cf. Goossens 2005, Tutin *et al.* 2006). La première dimension (« expérientiation ») réunit des collocatifs exprimant, sans spécification supplémentaire, qu'une personne expérimente un affect (*ressentir de la colère*). La dimension « aspect » concerne la manière dont un affect se déroule dans le temps. Elle regroupe des collocatifs indiquant le caractère ponctuel (*instant de stupeur*) ou duratif de l'affect (*climat d'euphorie*), le début (*être submergé par le chagrin*), la fin (*perdre son enthousiasme*), l'augmentation (*irritation grandissante*) ou la diminution (*la rage s'émousse*). On retrouve les mêmes sous-catégories aspectuelles pour la dimension « causativité » (*inspirer le mépris*, *dissiper l'amertume*, *décupler la rage*, *désamorcer la colère*), auxquelles s'ajoute une catégorie aspectuellement « neutre » (*causer de la surprise*). La dimension « contrôle » réunit les collocatifs dénotant l'(in)capacité de l'expérienceur à contrôler l'affect (*ne pas arriver à réprimer son agacement*). La dimension « manifestation » se réfère à la façon dont certains affects sont extériorisés par l'expérienceur : la manifestation « active » (*montrer son agacement*), celle qui est plutôt « subie » (*pleurer de rage*) ou « verbale » (*soupir d'admiration*) et la manifestation « externe » (*deviner la surprise de qqn*). La dimension « verbalisation » distingue les modalités « communicatif » (*avouer sa déception*) et « émotif » (*crier son indignation*). Enfin, la dimension « intensité » recense des collocatifs référant au degré d'intensité de

---

<sup>11</sup> Les cooccurrents sans relation sémantique et/ou syntaxique avec le pivot ainsi que les cooccurrents appartenant à la catégorie des mots grammaticaux ont été exclus du traitement.

l'affect (*respect profond*), et la dimension « polarité » concerne l'évaluation de l'affect sur deux axes : l'axe « interne » se réfère au caractère (dés)agréable de l'affect (*stupeur douloureuse*), alors que l'axe « externe » renvoie à des évaluations axiologiques du type bon / mauvais (*jalousie malade*).

Les résultats du codage sémantique ont été intégrés dans l'application *EmoGrammes* (cf. section 4.4), qui permet entre autres de calculer le degré de spécificité des dimensions sémantiques pour un pivot donné. Les dimensions codées peuvent ainsi servir de *tertium comparationis* lorsqu'il s'agit de contraster la signification de synonymes partiels à travers leur combinatoire lexicale. Ainsi, une analyse effectuée à l'aide d'*EmoGrammes* (cf. figure 8) permet de constater que les pivots *surprise* et *stupeur* s'opposent au niveau de la polarité exprimée par leurs collocatifs : alors que *surprise* se combine aussi bien avec des collocatifs dénotant le caractère agréable de l'affect (*bonne / heureuse* ~) qu'avec ceux indiquant son caractère désagréable (*mauvaise* ~), les collocatifs de *stupeur* réfèrent exclusivement au caractère désagréable de l'affect (~ *douloureuse / indignée*).

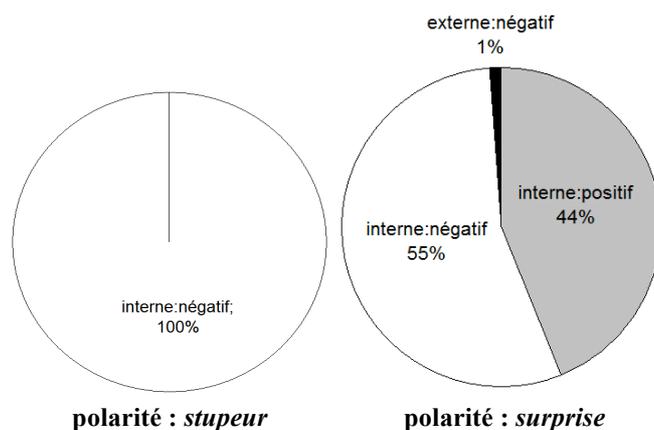


Figure 8. EmoGrammes avec la dimension de polarité et ses valeurs pour les lexies *stupeur* et *surprise*

Grâce à de multiples possibilités de paramétrage, les « émo-grammes » permettent des analyses à granularité variable qui

vont de la comparaison entre deux lexies jusqu'à l'opposition de champs lexicaux entiers, aussi bien à l'intérieur d'une même langue qu'au niveau interlinguistique (cf. Grutschus & Kern à par.) et rendent également possible l'étude de la distribution des dimensions à travers les champs sémantiques (Novakova & Melnikova 2013).

### 5.2 Grille syntaxique<sup>12</sup> pour l'analyse des lexies d'émotion

Les associations entre les mots-pivots (lexies d'émotion) et leurs collocatifs spécifiques sont également codées en fonction du type de relation syntaxique qu'ils entretiennent. La grille contient une liste de codes grammaticaux ainsi que leurs étiquettes péri-phrasées.

Par exemple, n31 : nom (+ préposition) + nom pivot : *moment de surprise* ; a30 : nom + adjectif pivot épithète : *un air surpris* ; v22 : verbe pivot + complément d'objet direct : *surprendre tout le monde*.

En guise d'illustration de la grille, nous présenterons une brève analyse des constructions syntaxiques privilégiées traduisant la dimension « aspect » dans les champs de SURPRISE et de DÉCEPTION (cf. la figure 9).

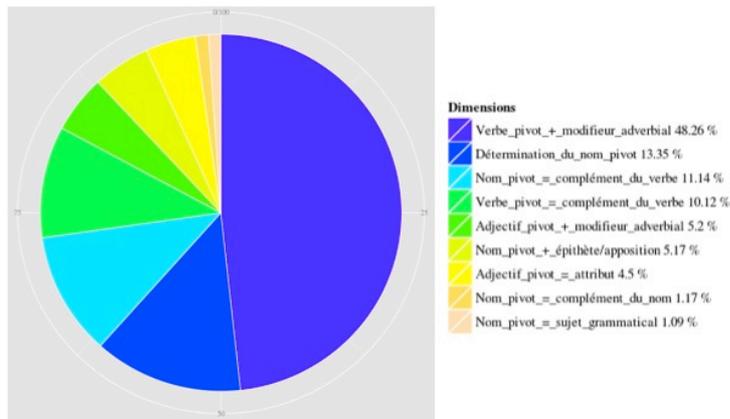


Figure 9. Les constructions syntaxiques spécifiques de l'« aspect »

<sup>12</sup> Cette grille a été élaborée à partir des travaux de l'équipe colonaise (Blumenthal 2007) et complétée par l'équipe EMOLEX.

La dimension « aspect » se présente ici comme une dimension transversale, concernant davantage les pivots verbaux et nominaux. Les verbes pivots sont le plus souvent modifiés par un adverbe, exprimant alors des valeurs ponctuelles et/ou itératives (*surprendre quelquefois, décevoir encore*). En complément d'un autre verbe, ils véhiculent en général des valeurs duratives (*continuer à surprendre, ne pas cesser d'étonner*). Les noms pivots apparaissent le plus souvent avec des déterminants nominaux ou adverbiaux, référant alors essentiellement à la valeur « ponctuel / itératif » (*un lot de surprises, bien des surprises*).

### **5.3 Les structures actanciennes des lexies des émotions**

Dans l'organisation pyramidale du traitement des données linguistiques, l'analyse des structures actanciennes fait le lien entre le niveau de l'analyse sémantique et syntaxique (sections 5.1 et 5.2) et le niveau de l'analyse discursive (section 5.4), en s'inscrivant dans une perspective « micro-discursive ». Ces structures ont été analysées pour les noms (Novakova *et al.* 2013b) et pour les verbes (Novakova *et al.* 2013a) de deux champs sémantiques : SURPRISE et RESPECT.

Dans la lignée de la *Role and Reference Grammar* (RRG) de Van Valin & LaPolla (1997) et de la *Théorie Sens-Texte* de Mel'čuk *et al.* (1995), nous distinguons, pour les verbes comme pour les noms :

- la valence syntaxique : les actants syntaxiques, régis par le verbe ou le nom, réalisés en surface et comptés dans la construction verbale ou nominale ;
- la valence sémantique : les actants sémantiques correspondant à des rôles clés comme l'expérienceur, l'objet ou la cause de l'affect. Ces actants sémantiques peuvent soit correspondre à des actants syntaxiques, régis par le verbe ou le nom (*Il estimait son courage*), soit à des ajouts (modificateurs du syntagme nominal – *Sa surprise en apprenant/quand elle apprit l'infidélité de Jean* –, compléments satellites non régis par le verbe – *Il nous a surpris avec cet incroyable spectacle*) qui n'ont pas le statut d'actants syntaxiques.

### *Traitement des lexies d'émotion*

Pour chacune des lexies des trois champs étudiés nous avons ainsi identifié, en analysant manuellement de nombreuses concordances :

- les combinaisons d'actants (rôles sémantiques codés conventionnellement X, Y, Z, etc.) attestées dans les corpus ;
- la fonction syntaxique des actants impliqués dans ces combinaisons (en utilisant la grille syntaxique présentée dans la section 5.2), et donc leur statut ou non d'actant syntaxique ;
- la forme des syntagmes associés à chaque actant (en spécifiant notamment les éventuelles prépositions introductives) ;
- le type d'emploi verbal : actif, passif, pronominal.

Nous avons procédé en deux étapes : une étape d'observation systématique (100 à 500 exemples sélectionnés aléatoirement pour chaque pivot) et une étape de recherche ciblée, pour vérifier et compléter les premières observations (utilisation de requêtes utilisant l'annotation syntaxique du corpus).

Nous avons ainsi décrit très finement le système actantiel tant nominal que verbal d'une portion importante du lexique des émotions. Nous avons pu identifier des actants véhiculant des rôles sémantiques moins décrits que les traditionnels expé-rienceur (X), cause (Z) ou objet (Y). Nous relevons ainsi des cas de fusion de deux actants, tels l'expérienceur (X) et l'objet (Y) pour les verbes pronominaux (*Mais les deux hommes (X+Y) s'estiment.* Le Figaro), ou de dédoublement d'un actant, en l'occurrence la cause (*Le daphné (Z) a surpris les néophytes par son odeur acidulée* (Zinstr). Ouest-France).

Toutes ces informations ont été implémentées dans un module de requête spécifique de l'application *EmoLing*, qui permet de sélectionner les informations à afficher en restreignant la langue, le champ sémantique, les lemmes, les catégories, les fonctions syntaxiques, etc.

Lors de l'affichage des résultats de la requête il est possible de visualiser, pour chaque lexie, des exemples sélectionnés dans les corpus.

**Structures actancielles**

Afficher 10 lignes Search:

Langue	Classe	Lemme	Exemples	X	Y	Z	X+Y	Y+Z	Zinstr	X+Z
Français	SURPRISE	surprise_N	Exemples	<ul style="list-style-type: none"> <li>chez SN</li> <li>dans SN</li> <li>de SN</li> <li>det poss</li> <li>adj relationnel</li> </ul>	-	<ul style="list-style-type: none"> <li>quandP</li> <li>pprésent</li> <li>de SN</li> <li>devant SN</li> <li>à Vinf</li> <li>de Vinf</li> </ul>	-	-	-	-
Français	SURPRISE	étonnement_N	Exemples	<ul style="list-style-type: none"> <li>chez SN</li> <li>dans SN</li> <li>de SN</li> <li>det poss</li> <li>adj relationnel</li> </ul>	-	<ul style="list-style-type: none"> <li>quandP</li> <li>pprésent</li> <li>de SN</li> <li>devant SN</li> <li>à Vinf</li> <li>de Vinf</li> <li>queP</li> </ul>	-	-	-	

Figure 10. Affichage des structures actancielles des lexies sélectionnées

L'analyse linguistique se poursuit, au niveau transphrastique, avec l'étude des positions occupées par les lexies d'émotion dans la macro-structure discursive (ou textuelle).

#### 5.4 Analyse textuelle

Selon la théorie du *Lexical Priming* de Hoey (2005), les mots ont des préférences ou des aversions pour certaines positions et, de là, pour certaines fonctions grammaticales (« colligations ») au niveau syntagmatique, phrastique, au sein du paragraphe ou du texte (« colligations textuelles »). Le module *EmoConc* permet de repérer et d'extraire de manière automatique les colligations, et plus précisément, la place du pivot dans la phrase, le paragraphe et le texte divisés en différentes tranches (titre, position initiale, médiane ou finale). Cette fonctionnalité a été appliquée à l'analyse textuelle des lexies des émotions (Novakova & Sorba 2013, 2014). Il a été ainsi démontré, par exemple, que *stupéur* et *jalousie* génèrent des structurations textuelles différentes en fonction de leur sémantisme: *stupéur*, affect réactif causé, ponctuel, intensif se rencontre de préférence dans des positions initiales (titre, début de texte, du paragraphe ou de la phrase), ce qui génère plus d'attentes chez le lecteur, et jamais en position finale dans un texte. *Jalousie*, affect interpersonnel, duratif, négatif, moins intensif, est très largement attestée en milieu de texte (plus des deux tiers des occurrences) ; la lexie génère plus d'associations lexicales avec d'autres mots d'émo-

tions et crée moins d'attentes chez le lecteur. Ainsi le scénario que les deux lexies engendrent dans la séquence textuelle est différent et tributaire de leurs propriétés sémantiques spécifiques. En bref, le mot à lui seul permet de « télécommander » la structuration d'un paragraphe ou d'un texte (Blumenthal 2014).

## 6. Conclusion

Bien qu'*EmoBase* ait été construite pour répondre aux besoins d'une analyse contextualisée du lexique des émotions dans cinq langues (français, anglais, allemand, espagnol et russe), elle peut servir à l'étude de n'importe quel autre type de vocabulaire. Un travail important a été réalisé pour la collecte, le traitement des corpus (balisage XML, annotation morpho-syntaxique et analyse en dépendances), ainsi que pour la conception d'outils d'extraction des mots et de leurs associations lexicales. Une méthode d'extraction novatrice des expressions récurrentes de longueur *n* autour d'un pivot donné a été développée (Kraif *et al.* 2014) et implémentée dans le module *EmoConc* d'*EmoBase*.

La méthodologie de sélection et d'analyse des lexies d'émotion peut être aussi appliquée à l'étude d'autres types de lexique. L'accès à la ressource et à ses corpus comparables et parallèles dans les cinq langues est simple et rapide et se fait via la création d'un compte personnel (formulaire à remplir à la page d'accueil d'*EmoBase*), qui permet de profiter de toutes les fonctionnalités de la base et de ses corpus. On peut également se connecter en tant que visiteur, statut qui permet d'utiliser la ressource (visualisation des lexicogrammes, *EmoGrammes* etc.) sans avoir accès aux concordances. Les guides d'utilisation sont disponibles en format pdf dans chaque module de la base. En un an, depuis sa mise en ligne en juin 2013, *EmoBase* a été visitée par plus de 5 000 utilisateurs de 20 pays (Allemagne, Tunisie, Chypre, Espagne, Belgique, France, Russie etc.).

## Références bibliographiques

- Aït-Mokhtar S. *et al.* (2002). « Robustness beyond Shallowness: Incremental Deep Parsing », *Natural Language Engineering* 8 : 121-144.

- Anscombre J.-C. (1995). « Morphologie et représentation événementielle : le cas des noms de sentiment et d'attitude », *Langue Française* 105 : 40-54.
- Attardi G. et al. (2007). « Multilingual Dependency Parsing and Domain Adaptation using DeSR », in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague : ACL, 1112-1118.
- Blumenthal P. (2007). « A Usage-based French Dictionary of Collocations », in Kawaguchi Y. et al. (éd.) *Corpus-Based Perspectives in Linguistics*. Amsterdam : Benjamins, 67-83.
- Blumenthal P. (2014). « Caractéristiques et effets de la complexité sémantique des noms d'affects », in P. Blumenthal et al. (éd.) *Les émotions dans le discours. Emotions in discourse*. Frankfurt s. M. : Lang, 175-186.
- Blumenthal P. et al. (2005). « Kombinatorische Profile und Profilkontraste : Berechnungsverfahren und Anwendungen », *Zeitschrift für romanische Philologie* 121, 1, 49-83.
- Buvet P.-A. et al. (2005). « Les prédicats d'affect », *LIDIL* 32 : 123-143.
- Charest S. et al. (2010). « Au-delà de la paire de mots : extraction de cooccurrences syntaxiques multi-lexémiques », *Actes de TALN 2010*, Montréal, 19-23 juillet 2010, s. p [en ligne : [http://iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010\\_submission\\_90.pdf](http://iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_90.pdf)].
- Diwersy S. (2012). *Kookkurrenz, Kontrast, Profil. Korpusinduzierte Studien zur lexikalisch-syntaktischen Kombinatorik französischer Substantive (mit ergänzenden Betrachtungen zum Deutschen)*. Berlin et al. : de Gruyter.
- Diwersy S. & Kraif O. (2013). « Observations statistiques de co-occurents lexico-syntaxiques pour la catégorisation sémantique d'un champ lexical » in F. Baider & G. Cislaru (éd.) *Cartographie des émotions. Propositions linguistiques et sociolinguistiques*. Paris : PSN, 55-69.
- Ducrot O. (1975). « Je trouve que », *Semantikos* 1, 1 : 63-88.

- Evert S. (2009). « Corpora and collocations », in A. Lüdeling & M. Kytö (éd.) *Corpus Linguistics. An International Handbook*, vol. 2. Berlin : Mouton de Gruyter, 1212-1248.
- Firth J. R. (1957). *Papers in Linguistics*. London : Oxford University Press.
- Goossens V. (2005). « Les noms de sentiment : esquisse de typologie sémantique fondée sur les collocations verbales », *Lidil* 32 : 103-121.
- Goossens V. (2011). Propositions pour le traitement de la polysémie régulière des noms d'affect. Thèse de doctorat, Grenoble : Université Stendhal.
- Grutschus A. & Kern B. (à par.). « *Decepción, surprise, colère et furia* : exploration d'une méthode statistique en lexicologie », *Zeitschrift für romanische Philologie* 130, 3.
- Grutschus A. et al. (2013). « La polarité des lexies des émotions : perspective combinatoire et contrastive », in F. Baider & G. Cislaru (éd.) *Cartographie des émotions. Propositions linguistiques et sociolinguistiques*. Paris : PSN, 85-96.
- Hausmann F. & Blumenthal P. (2006). « Présentation : collocation, corpus, dictionnaires », *Langue française* 150 : 3-13.
- Heiden S. & Tournier M. (2000). « Lexicométrie textuelle, sens et stratégie discursive », in J. J. de Bustos Tovar (éd.) *Lengua, discurso, texto : Simposio internacional de análisis del discurso*. Madrid : Visor, 2287-2300.
- Hoey M. (2005). *Lexical Priming. A New Theory of Words and Language*. London/New York : Routledge.
- Keller R. (1995). *Zeichentheorie: Zu einer Theorie semiotischen Wissens*. Tübingen/Basel: Francke/UTB.
- Kern B. & Grutschus A. (2014). « *Surprise vs étonnement* : comportement discursif et perspectives contrastives » in P. Blumenthal et al. (éd.) *Les émotions dans le discours. Emotions in Discourse*. Frankfurt s. M. : Lang, 187-198.
- Kraif, O. (2008). « Comment allier la puissance du TAL et la simplicité d'utilisation ? L'exemple du concordancier bilingue ConcQuest », in *JADT 2008*, vol. 2. Lyon : PUL, 625-634.

- Kraif, O. *et al.* (2014). « Extraction de pivots complexes pour l'exploration de la combinatoire du lexique : une étude dans le champ des noms d'affect », in *Actes du 4<sup>e</sup> Congrès Mondial de Linguistique Française*. Berlin, juillet 2014, SHS Web of Conferences, 2663-2674.
- Mel'čuk I. *et al.* (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve : Duculot.
- Nivre J. *et al.* (2008). « Parsing the SYNTAGRUS Treebank of Russian », in *Proceedings of the 22<sup>nd</sup> International Conference on Computational Linguistics (Coling 2008)*, Manchester, August 2008, 641-648.
- Novakova I. *et al.* (2013a). « Interactions entre profil discursif et structures actanciennes : l'exemple des verbes de surprise et de respect », *Langue Française* 180 : 31-46.
- Novakova I. *et al.* (2013b). « Interactions entre profil discursif et structures actanciennes : l'exemple des noms de surprise et de respect » in F. Baider & G. Cislaru (éd.) *Cartographie des émotions. Propositions linguistiques et sociolinguistiques*. Paris : PSN, 71-84.
- Novakova I. & Melnikova E. (2013). « Vers un modèle fonctionnel pour l'analyse du lexique des émotions dans cinq langues européennes », *Bulletin de la Société de linguistique de Paris* CVIII, 1 : 131-160.
- Novakova I. & Sorba J. (2013). « *Stupéfier* et *jalouser* dans les séquences textuelles journalistiques : quel profil discursif pour quelle stratégie argumentative ? », *Le discours et la langue* 4, 1 : 203-220.
- Novakova I. & Sorba J. (2014). « L'émotion dans le discours. la recherche du profil discursif de *stupeur* et de *jalousie* » in P. Blumenthal *et al.* (éd.) *Les émotions dans le discours. Emotions in Discourse*. Frankfurt s. M. : Lang, 161-175.
- Novakova I. & Tutin A. (éd.) (2009). *Le lexique des émotions*. Grenoble : ELLUG.
- Sinclair J. M. (1991). *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

*Traitement des lexies d'émotion*

- Tutin A. *et al.* (2006). « Esquisse de typologie des noms d'affect à partir de leurs propriétés combinatoires », *Langue française* 150 : 45-55.
- Van Valin R. D. & LaPolla R. J. (1997). *Syntax : Structure, Meaning, Function*. Cambridge : Cambridge University Press.