Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles
Lille, France, 28 juin au 2 juillet 2021

Une étude des avis en ligne: généralisabilité d'un modèle d'évaluation

Hyun Jung Kang Iris Eshkol-Taravella
MoDyCo UMR7114, 200 Avenue de la République, 92001 Nanterre, France
clinguist.hjkang@gmail.com, ieshkolt@parisnanterre.fr

RÉSUMÉ

Ce travail se situe dans la continuité de nos travaux antérieurs proposant le modèle d'évaluation portant sur des avis en ligne sur des restaurants. Le modèle est composé de quatre catégories : l'opinion (positive, négative, mixte), la suggestion, l'intention et la description. Cet article vise à tester la généralisabilité du modèle en l'appliquant sur deux corpus supplémentaires : un corpus relevant d'un autre domaine (celui de l'hôtellerie) et un corpus écrit dans une autre langue (le coréen). Nous avons présenté l'annotation manuelle et la détection automatique de ces catégories en nous appuyant sur différents modèles de l'apprentissage de surface (SVM) et l'apprentissage profond (LSTM).

ABSTRACT

A Study of Online Reviews : Generalizability of the Evaluation Model

This work is a continuation of our previous work proposing a model of evaluation on online restaurant reviews. The model comprises four categories : opinion (positive, negative, mixed), suggestion, intention, and description. This article attempts to test the generalizability of the proposed model on two other corpora : a corpus from another domain (hotel) and a corpus written in another language (Korean). We presented manual annotation and automatic detection of these categories based on traditional machine learning (SVM) and deep learning (LSTM).

MOTS-CLÉS : fouille d'opinion, avis en ligne, classification supervisée, généralisabilité, coréen.

KEYWORDS: opinion mining, online reviews, supervised learning, generalizability, Korean.

1 Introduction

Dans un monde interconnecté confronté à l'augmentation exponentielle du volume de données, de nombreux internautes se réfèrent au bouche-à-oreille électronique (eWOM) d'inconnus à travers diverses plateformes (blogs, forums ou sites dédiés à la critique) (Wachsmuth *et al.*, 2014). Ainsi, les évaluations diffusées en ligne ont un pouvoir conséquent car elles influencent la prise de décision des internautes. En traitement automatique des langues (TAL), elles sont concernées par la tâche de la fouille d'opinions. Cependant, à la différence de nombreux travaux dans ce domaine, nous dépasserons largement la notion d'opinion positive ou négative. Dans cette optique, nous nous interrogerons sur la manière dont une évaluation émerge dans le langage en tant que jugement axiologique. L'évaluation renvoie au second degré de la subjectivité (Kerbrat-Orecchioni, 1980), à laquelle l'axiologie positive ou négative du locuteur est associée. Elle fait intervenir un ensemble de normes et de valeurs qui s'inscrit dans le cadre des perceptions collectives, sensibles aux différences sociales et culturelles, et qui évolue au fil du temps.

Dans ce cadre, nous avons proposé un modèle d'évaluation fondé sur l'observation manuelle du corpus d'avis postés en ligne sur des restaurants (Eshkol-Taravella & Kang, 2019; Kang & Eshkol-Taravella, 2020), que nous appellerons « RestoFR [1] ». Le modèle décrit différents types d'évaluation, cette dernière se composant de quatre catégories : l'opinion (positive/négative/mixte), la suggestion, l'intention et la description. L'opinion représente l'évaluation de la valeur du restaurant dans une dimension axiologique qui comporte les polarités positive, négative et mixte (e.g., les adjectifs évaluatifs, les lexiques des émotions ou des sentiments et les modifieurs). La suggestion vise d'une part à améliorer les produits (ou services) du restaurant et d'autre part à donner des conseils aux autres consommateurs (e.g., les verbes de parole, le mode impératif et le mode conditionnel). L'intention renvoie au souhait du client de renouveler ou non son expérience dans le restaurant, permettant aux restaurateurs de se renseigner sur les actions futures qu'un client a l'intention d'engager (e.g., les verbes au futur et le préfixe verbal d'itération « re- »). La description concerne les informations factuelles associées à l'expérience vécue, qui révèle aux lecteurs l'arrière-plan de cette dernière. Nous avons présenté la détection automatique de ces catégories fondée sur différents modèles de l'apprentissage de surface (Naïve Bayes, SVM, Logistic Regression) et l'apprentissage profond (CNN, LSTM) (Eshkol-Taravella & Kang, 2019; Kang & Eshkol-Taravella, 2020). La meilleure F-mesure a été obtenue grâce au classifieur SVM, donnant un score de 0,88 [2].

L'une des étapes essentielles de l'élaboration d'un modèle consiste à évaluer sa généralisabilité (Clark & Watson, 2016). La possibilité de généraliser est essentielle dans le contexte de la mondialisation, et particulièrement de l'ère numérique, dans lequel nous ne pouvons plus nous limiter aux données textuelles d'un seul domaine ou d'une seule langue. Dans cet article, nous avons pour objectif de vérifier si le modèle élaboré pour le corpus RestoFR peut s'appliquer à un corpus relevant d'un autre domaine (celui de l'hôtellerie) et à un corpus écrit dans une autre langue (le coréen). Si notre approche est valable, les six catégories d'évaluation devraient également être identifiées au sein des nouveaux corpus. L'annotation manuelle et la détection automatique réalisées sur les avis concernant des hôtels (HotelFR) sont décrites dans la seconde section. En ce qui concerne le corpus coréen (RestoKR), présenté dans la troisième section, nous nous en sommes tenues à son annotation manuelle car le prétraitement de cette langue est complexe et diffère largement de celui du français.

2 Application du modèle au corpus portant sur l'hôtellerie

La réussite de la généralisation d'une approche dans un autre domaine dépend principalement de deux éléments (Szarvas *et al.*, 2012) : d'abord, les domaines source et cible doivent être étroitement liés pour permettre le partage des connaissances ; deuxièmement, l'adaptation doit se fondre sur les points communs des deux domaines, tout en garantissant les caractéristiques particulières du domaine cible. Bien que l'hôtellerie et la restauration utilisent des lexiques différents, il s'agit dans les deux domaines de situations vécues dans un lieu, dont l'évaluation porte sur le processus de satisfaction et la valorisation de l'expérience. Trois hôtels situés à Paris ont été choisis au hasard, et pour chacun de ceux-ci, nous avons extrait d'un site internet [3] 99 avis rédigés en français. Les avis ont été extraits selon leur date de parution : ceux qui sont antérieurs au mois de février 2020 ont été collectés, dans la

1. Il s'agit de 6 287 avis collectés sur un site internet (https://www.lafourchette.com), correspondant à 17 268 phrases, avec une moyenne de dix mots par phrase.

2. Le nombre limité de pages ne permet pas de décrire le modèle et les traits linguistiques de manière plus détaillés. Nous pouvons mettre à votre disposition les articles concernés.

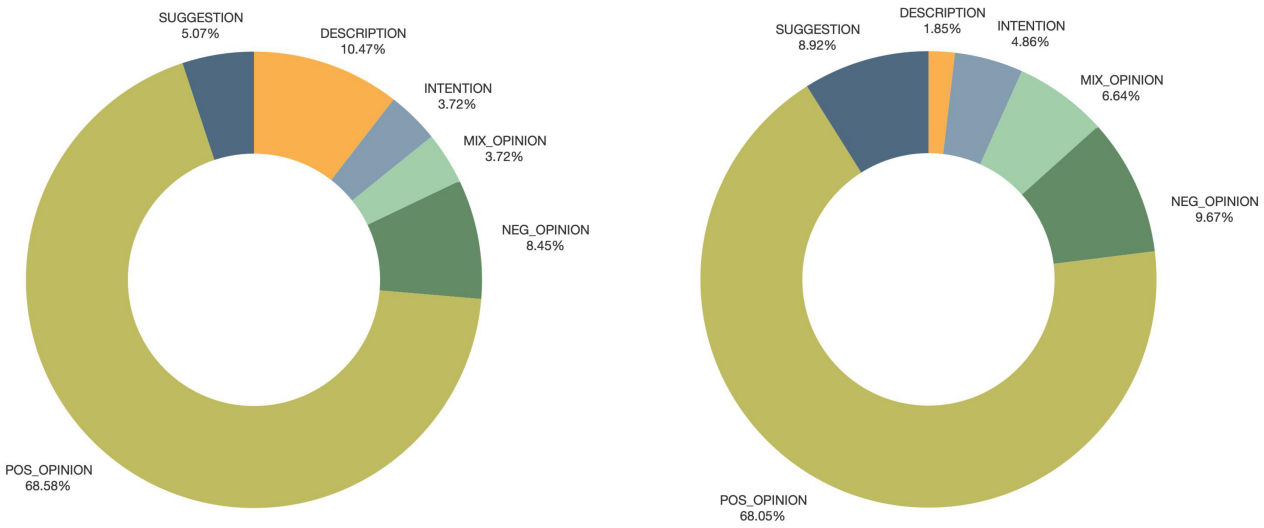
3. https://www.tripadvisor.com [consulté le 3er janvier 2021].

limite de 33 avis. Ils ont été segmentés en phrases en fonction des signes de ponctuation, produisant un total de 296 phrases (HotelFR). Le corpus a été annoté manuellement selon la typologie d'évaluation élaborée (POS_OPINION, NEG_OPINION, MIX_OPINION, SUGGESTION, INTENTION et DESCRIPTION), qui a ensuite été utilisée pour sa détection automatique.

2.1 Annotation manuelle

La tâche d'annotation a été réalisée par deux doctorantes en linguistique. Ces dernières ont suivi le guide d'annotation qui préconise d'attribuer une étiquette à chaque phrase. En cas d'ambiguïté, lorsque plusieurs catégories étaient pertinentes, les catégories les moins représentées dans le corpus (l'intention ou la suggestion) ont été retenues [4]. Pour valider la typologie d'évaluation et évaluer sa généralisabilité, l'accord inter-annotateur entre ces deux annotateurs est calculé en appliquant la mesure Kappa de Cohen (Cohen, 1960). Nous avons obtenu un accord inter-annotateur (AIA) de 0,83, considéré comme 'presque parfait' selon l'échelle de Landis & Koch (1977). Ce score montre que les catégories et les règles d'annotations sont bien définies et expliquées dans le guide d'annotation et que la typologie proposée peut être généralisée à d'autres domaines.

Néanmoins, l'observation du corpus a permis de révéler une différence significative par rapport à RestoFR. La répartition des catégories dans ce corpus, présentée dans la figure 1a, s'est avérée non homogène. En comparaison de RestoFR (voir figure 1b), la proportion de description s'est révélée être significativement plus importante, passant de 1,85 % à 10,47 %. Cette croissance est due aux particularités de l'hôtellerie, secteur dans lequel le confort est un aspect primordial ; les équipements, les installations et la localisation sont des éléments qui permettent de le garantir, qui font partie dans notre typologie d'une catégorie de description.



(a) HotelFR (b) RestoFR

FIGURE 1: Répartition des catégories

4. Ce choix a été fait dans la lignée de certains travaux de psychologie (Carreiras *et al.*, 1995; Gernsbacher, 1990; Kim *et al.*, 2004) affirmant que le contenu présenté au début de la phrase semble avoir un effet plus important dans la détermination du jugement final.

2.2 Annotation automatique

Nous avons considéré trois corpus d'entraînement différents : H (67 % de HotelFR de manière stratifiée), R (l'ensemble de RestoFR) et H+R (l'ensemble de RestoFR et 67 % de HotelFR de manière stratifiée). L'évaluation a été effectuée sur le corpus de test qui représente 33 % de HotelFR de manière stratifiée (98 phrases). En règle générale, nous avons adopté la méthodologie exploitée précédemment dans auteur1 & auteur2 (2019, 2020), dont nous ne présentons ici que la synthèse.

Nettoyage et prétraitement. La procédure comprend les étapes suivantes : le passage des mots en minuscules, le remplacement des chiffres par « NUM », le remplacement des émoticônes par « emoPOS » ou « emoNEG » selon leur polarité, la normalisation des abréviations (resto en restaurant, par exemple) et la lemmatisation au moyen de StanfordCoreNLP [5].

Classification. Parmi les divers algorithmes, nous avons choisi ceux ayant produit les meilleurs résultats sur le corpus RestoFR pour l'apprentissage de surface et l'apprentissage profond (SVM et LSTM respectivement). Le classifieur SVM a été appliqué à l'aide de la bibliothèque scikit-learn [6] (Pedregosa *et al.*, 2011) et le second avec les bibliothèques Keras (Chollet *et al.*, 2015) et TensorFlow. En ce qui concerne le classifieur SVM, les données textuelles ont d'abord été représentées selon la méthode de représentation vectorielle (`CountVectorizer` ou `TfidfVectorizer`) et à l'aide du paramètre n_gram. Leur meilleure combinaison a été obtenue grâce à une procédure de grille de recherche (`GridSearch`). Les caractéristiques prises en compte lors de l'apprentissage sont : les catégories morphosyntaxiques jugées pertinentes proposées par StanfordCoreNLP, les différentes variations des verbes, la négation, la conjonction mais, les mots positifs et négatifs, les scores de polarité et de subjectivité obtenus avec TextBlob [7], la position de la phrase dans l'avis, le nombre de caractères, la longueur de la phrase, la diversité et la densité lexicales, les ponctuations multiples et l'unité monétaire. Afin d'exploiter la technique des LSTM, nous avons pris en entrée une matrice d'embedding à l'aide de Word2vec. Cette dernière a été créée à nouveau pour chaque jeu d'entraînement. Par la suite, nous avons appliqué une couche avec 100 unités, envoyée à une couche dense avant une activation softmax. Les hyperparamètres choisis sont l'optimiseur Adam et une perte d'entropie, et la taille du batch était de cinq, avec sept époques. Pour toutes les expériences, une validation croisée stratifiée à cinq plis a été effectuée et le paramètre `class_weight` au mode équilibré (`'balanced'`) a été appliqué afin de résoudre le problème de la disproportion des classes.

Résultats. Pour chaque jeu de données d'entraînement, la moyenne pondérée de la précision, du rappel et de la F-mesure a été calculée lorsque SVM et LSTM ont été appliqués (tableau 1). Le classifieur SVM entraîné avec H+R donne le meilleur score, dont la F-mesure est de 0,8064. Grâce à la combinaison des données (HotelFR et RestoFR), les performances de SVM et LSTM se sont considérablement améliorées.

5. Stanford CoreNLP, https://stanfordnlp.github.io/CoreNLP/download.html [consulté le 3 janvier 2021].

6. http://scikit-learn.org/stable/ [consulté le 3 janvier 2021].

7. Une bibliothèque pour le traitement des données textuelles (https://textblob.readthedocs.io/en/dev/index.html).

TABLE 1: Moyenne pondérée de la F-mesure (HotelFR)

Jeu d'entraînement	**F-mesure**	
	SVM	**LSTM**
H	0,6366	0,0189
R	0,7992	0,5160
H+R	**0,8064**	**0,6684**

D'après une observation manuelle du corpus, il existe des moyens systématiques relativement indépendants du domaine pour exprimer une évaluation. Grâce à des constructions linguistiques communes, les données de RestoFR semblent avoir comblé le manque d'informations de HotelFR, malgré le fait que les deux ensembles de données proviennent de domaines différents (restaurants et hôtels). Ainsi, lorsque nous ne disposons pas de suffisamment de données annotées pour un domaine en particulier, celles relatives à un domaine qui s'en rapproche peuvent pallier ce déficit et donc permettre de réduire les coûts d'annotation nécessaires pour couvrir le manque des données.

TABLE 2: Précision, rappel et F-mesure de chaque catégorie (H+R et SVM)

	POS_OPINION	NEG_OPINION	MIX_OPINION	SUGGESTION	INTENTION	DESCRIPTION
Précision	0,84	0,67	0,67	1,00	1,00	0,60
Rappel	0,97	0,44	0,50	0,80	1,00	0,30
F-mesure	0,90	0,53	0,57	0,89	1,00	0,40

La précision et le rappel du meilleur résultat (H+R avec SVM) pour chaque catégorie sont présentés dans le tableau 2. L'intention est détectée avec la meilleure performance (la F-mesure étant 1,00), suivie par l'opinion positive et la suggestion dont la F-mesure se situe autour de 0,90. En revanche, la description possède la plus mauvaise performance (0,40). Ce résultat s'explique en partie par le manque d'échantillons des descriptions et donc par le faible nombre de caractéristiques fournies durant l'entraînement. De plus, cette catégorie est très hétérogène et varie en fonction du profil du client, ce qui cause l'apparition d'un large éventail de vocabulaires et de contextes. Ainsi, notre typologie d'évaluation peut être généralisée à d'autres domaines tant que l'évaluation porte sur des expériences à un endroit donné (un lieu touristique ou un magasin par exemple).

3 Application du modèle sur le corpus en langue coréenne

L'évaluation, qui émerge dans le langage en tant que jugement axiologique, peut être perçue et interprétée de différentes manières selon la culture et la langue. Plus précisément, l'interprétation axiologique de la valeur – ce qui est bon ou mauvais, agréable ou désagréable, satisfaisant ou insatisfaisant, souhaitable ou à éviter – peut varier d'une culture à l'autre. Cette section a donc pour but d'évaluer la généralisabilité de notre modèle d'évaluation proposé sur une autre langue, le coréen (RestoKR). Nous avons choisi six restaurants qui se trouvent à Séoul et proposent de la nourriture coréenne. Les avis sur les restaurants ont été sélectionnés en fonction de leur date de parution : pour

chaque restaurant, nous avons extrait dix des avis les plus récents du mois de mai 2020. Le corpus coréen est donc constitué de 60 avis (246 phrases), collectés sur le même site que le corpus HotelFR. Bien que les avis soient rédigés en coréen, ils sont comparables à ceux de RestoFR puisqu'il s'agit des évaluations effectuées par les clients. Les études coréennes sur la fouille d'opinions exploitent différentes méthodes de classification (l'apprentissage non supervisé, l'apprentissage de surface et l'apprentissage profond). Cependant, en raison de la spécificité du coréen en tant que langue agglutinante, de nombreuses études adoptent des méthodes particulières s'appuyant non sur des mots (comme pour le français) mais sur d'autres unités comme les morphèmes ou les jamos [8]. Ainsi, le prétraitement et la détection automatique sont complexes et bien différents de ceux du français. Dans cette étude, nous nous sommes donc limités à son annotation manuelle.

3.1 Annotation manuelle

Deux annotatrices coréennes (une doctorante en linguistique et une autre en littérature) ont effectué l'annotation manuelle du corpus RestoKR. La tâche d'annotation a consisté à attribuer à chaque phrase une étiquette parmi des catégories prédéfinies. Selon la mesure Kappa de Cohen (Cohen, 1960), nous avons obtenu un AIA de 0,92, un accord considéré comme « presque parfait » (Landis & Koch, 1977). Ce score a ainsi permis d'appuyer la reproductibilité de notre typologie d'évaluation sur une autre langue. La figure 2 illustre la répartition des catégories, qui a permis de mettre en lumière les spécificités du corpus coréen.

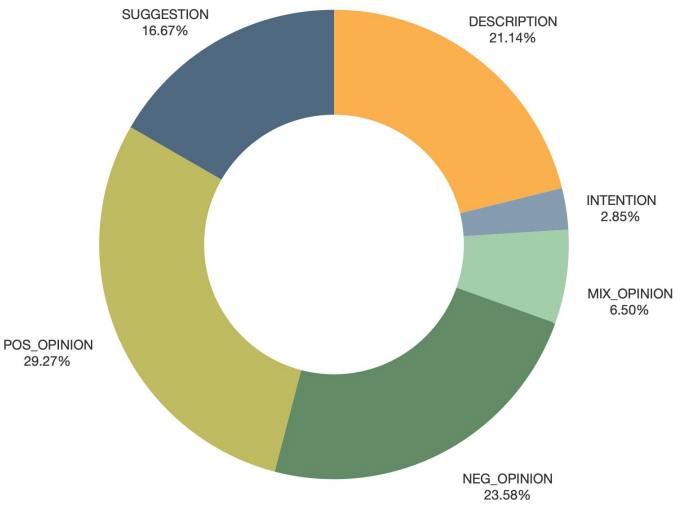


FIGURE 2: Répartition des catégories d'évaluation (RestoKR)

En premier lieu, la répartition des catégories est plus équilibrée que celle des corpus français (HotelFR, RestoFR) (cf. figures 1a, 1b). L'opinion positive constitue toujours la classe majoritaire (29,27 %), mais de façon moins marquée que dans RestoFR (68,05 %). À l'inverse, l'opinion négative atteint 23,58 %, contre 9,5 % pour le corpus RestoFR. Bien que Jurafsky *et al.* (2014) aient mis en évidence le principe de pollyanna – une tendance générale à préférer les informations positives – dans leur

8. Un ensemble d'unités phonétiques de base représentant les consonnes et les voyelles de la langue coréenne.

corpus anglais, il semble que celui-ci ne s'applique pas à notre corpus coréen [9].

En outre, les annotatrices ont rencontré une difficulté liée à l'interprétation axiologique, qui peut varier selon le locuteur. Il s'agit de la distinction entre les catégories de description et d'opinion (positive/négative). À titre d'exemple, la phrase « 고기만두라고 하지만, 두부와 야채를 많이 넣은 만두에요 (Ce sont des raviolis au bœuf, mais ils sont remplis de beaucoup de tofu et de légumes) » a été considérée comme relevant de la description par la première annotatrice (A1), qui a estimé qu'il s'agissait simplement d'un inventaire des farces des raviolis. En revanche, la seconde annotatrice (A2) a relié cette phrase à l'opinion négative, considérant que les raviolis au bœuf devaient être principalement farcis de viande et non de tofu ou de légumes. Cette tendance a également été observée dans le corpus HotelFR, dont la proportion de description était élevée par rapport à RestoFR (figures 1a, 1b). La distinction de ces catégories peut ainsi s'avérer quelque peu délicate, car si certains lecteurs acceptent la description pour ce qu'elle est, d'autres vont plus loin et la perçoivent comme reflétant un état favorable ou non d'après eux (l'opinion). Cela montre que l'évaluation dépend d'un système de normes et de valeurs propre à chaque locuteur. Ainsi, bien que chaque langue emploie des indices linguistiques différents pour exprimer leur point de vue axiologique, la typologie d'évaluation que nous avons élaborée montre qu'elle représente assez fidèlement ce en quoi consiste l'évaluation.

4 Conclusion

Notre conclusion principale est que la typologie que nous avons élaborée en nous appuyant sur le corpus RestoFR peut être appliquée à la fois à un autre domaine proche concernant l'expérience dans un lieu (l'hôtel) et à une autre langue (coréen). Cependant, la distinction entre l'opinion et la description présente des difficultés, car l'interprétation de ces deux catégories peut varier selon le locuteur. En outre, la détection automatique de HotelFR montre que des données annotées pour un domaine particulier peuvent s'appliquer à d'autres domaines comparables disposant de peu de données, ce qui permet de réduire les coûts d'annotation nécessaires pour couvrir le manque de données. En perspective, il serait intéressant de réaliser une annotation (et des modèles de classification) multilabels afin de résoudre les cas où les deux catégories d'évaluation se superposent. Par ailleurs, il serait avantageux d'augmenter la taille du corpus de référence de différents domaines et langues, et notamment de réaliser la détection automatique en langue coréenne en utilisant les modèles de langue pré-entraînés sur différentes langues comme XLM-RoBERTa (Conneau *et al.*, 2019).

Références

CARREIRAS M., GERNSBACHER M. A. & VILLA V. (1995). The advantage of first mention in spanish. *Psychonomic Bulletin & Review*, **2**(1), 124–129.

CHOLLET F. *et al.* (2015). Keras. https://github.com/fchollet/keras.

CLARK L. A. & WATSON D. (2016). Constructing validity : Basic issues in objective scale development. *Psychological Assessment*, **7**(3), 309–319.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, **20**(1), 37–46.

9. Nous reconnaissons néanmoins que ce résultat peut être dû à la petite taille du corpus.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMÁN F., GRAVE E., OTT M., ZETTLEMOYER L. & STOYANOV V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv :1911.02116*.

ESHKOL-TARAVELLA I. & KANG H. J. (2019). Observation de l'expérience client dans les restaurants (Mapping Reviewers' Experience in Restaurants). In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN)-PFIA 2019-Volume II : Articles courts*, p. 361–370 : ATALA.

GERNSBACHER M. A. (1990). *Language Comprehension as Structure Building*. Psychology Press.

JURAFSKY D., CHAHUNEAU V., ROUTLEDGE B. & SMITH N. (2014). Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, **19**.

KANG H. J. & ESHKOL-TARAVELLA I. (2020). Les avis sur les restaurants à l'épreuve de l'apprentissage automatique (An Empirical Examination of Online Restaurant Reviews). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, p. 249–257 : ATALA.

KERBRAT-ORECCHIONI C. (1980). *L'énonciation de la subjectivité dans le langage (4e édition) [version Kindle iOS]*. Armand Colin.

KIM S.-I., LEE J.-H. & GERNSBACHER M. A. (2004). The advantage of first mention in korean the temporal contributions of syntactic, semantic, and pragmatic factors. *Journal of psycholinguistic research*, **33**(6), 475–491.

LANDIS J. R. & KOCH G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**(1), 159–174.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V. *et al.* (2011). Scikit-learn : Machine learning in python. *Journal of Machine Learning Research*, **12**, 2825–2830.

SZARVAS G., VINCZE V., FARKAS R., MÓRA G. & GUREVYCH I. (2012). Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, **38**(2), 335–367.

WACHSMUTH H., TRENKMANN M., STEIN B., ENGELS G. & PALAKARSKA T. (2014). A review corpus for argumentation analysis. In *International Conference on Intelligent Text Processing and Computational Linguistics*, p. 115–127 : Springer.