***Extraction of ingredient names from recipes by combining linguistic annotations and CRF selection*, T. Hamon & N. Grabar**

**Autres travaux mentionnés** : système Epicure (génération de description de recette grâce à des représentations en grammaires d’unification et ontologies)

**Objectif du travail** : détection et extraction automatique de noms d’ingrédients dans les recettes

**Procédé** : ressources lexicales et machine learning (CRF)

**Ressources utilisées** : liste d’ingrédients et d’aliments (cf. sources dans l’article), liste d’ustensiles, liste d’actions, liste de marqueurs de variantes, liste de stop-words, ressources pour la détection de quantités, ressource pour la détection de durées

**Corpus** : 13864 recettes pour l’entraînement, 2306 recettes pour l’évaluation (DEFT 2013 Challenge)

**Prétraitement** : POS tagging et lemmatisation avec Tree-Tagger (+ optionnel : analyse morpho-syntaxique avec FLEMM, ou YATEA pour les ressources lexicales)

**Première extraction avec système à base de règle** : module Perl Alvis ::TermTagger pour projeter ressources sur corpus et extraire les EN, identification des ingrédients les plus importants grâce au contexte et à leur position dans la recette, filtrage (inclusion lexicale, formes canoniques)

**Tri pour repérer ingrédients principaux parmi ceux extraits, avec ML :** Wapiti (CRF, algo L-BFGS)

***Information Extraction from Recipes*, R. Agarwal & K. Miller**

**Autres travaux mentionnés** : Converting IN-N-Out Orders into a Structured Form, de M. Gummelt & D. Brody, 2009

**Objectif du travail** : générer automatiquement une représentation de recettes interprétable par une machine en réduisant chaque étape de préparation à un prédicat (action) à un à deux arguments (ingrédient/patient, ustensile/agent)

**Procédé** : NER (Maximum Entropy Markov Model Classification) + SRL (=Semantic role labelling, algo à base de règles)

**Corpus** : très peu de données annotées (19 recettes pour le train, 12 pour le test), données semi-structurées en XML (balises pour les ingrédients au début, puis pour l’ensemble de la description de la préparation)

**Prétraitement** : Stanford pour pos-tagging (mais mauvais résultats car peu entraîné sur les verbes à l’impératif, qu’il tague souvent comme des noms)

**Résultats pour NER** : 7 catégories à détecter et 12 features données au classifieur (cf. articles pour le détail), mauvais résultats pour les groupes d’ingrédients et les ustensiles), très bons résultats pour les quantités et les unités de mesure

**Résultats pour SRL :** prédicats correspondent aux entités labellisées comme actions par le NER, SRL effectué par un système de règles (grâce à un arbre syntaxique) et un humain vérifie pour chaque prédicat si les bons arguments ont été trouvés. Résultats corrects (principales erreurs dues à erreurs de NER, les autres à une association d’un argument à un mauvais prédicat quand deux prédicats trop proches, mais une règles a été ajoutée, ou découpage d’un chunk en plusieurs arguments qd on a déjà plusieurs arguments au même). Module pour la résolution des co-références (mais fonctionnement non explicité clairement)