Guide d'annotation des expressions polylexicales pour le projet SimpleApprenant

Amalia Todirascu

Dans le cadre du projet SimpleApprenant, nous avons créé une base de données contenant des expressions polylexicales. La base a été alimentée à partir du Lexique-Grammaire (M.Gross, 1994, Laporte, 1988). La base contient des expressions polylexicales en grande majorité construites à l'aide d'un verbe. Elle contient des patrons syntaxiques d'utilisation de chaque expression. Nous avons reclassée manuellement les expressions en trois catégories : expressions idiomatiques, collocations, expressions figées. En effet, le public des apprenants est intéressé par les propriétés de chaque classe (sens figuré vs sens plus compositionnel et le figement syntaxique éventuel) que par une classification des expressions détaillée telle que présentée par (Kim et Baldwin, 2010).

Pour l'annotation de ces expressions, nous avons adopté une méthodologie inspirée par les pratiques réalisées dans le cadre du projet PARSEME, dans le cadre d'une annotation des expressions polylexicales verbales pour 22 langues. Dans le cadre de cette tâche, les expressions verbales ont été annotées, en particulier les expressions idiomatiques, les verbes à particule, les verbes réfléchis. En revanche, les collocations n'ont pas été annotées (dans le projet PARSEME on considère que les collocations sont seulement des associations de mots qui sont retrouvées fréquemment ensemble), ni d'autres expressions construites sur une tête nominale, adjectivale ou adverbiale.

Dans le cadre de notre projet, nous avons limité nos annotations à trois catégories d'expressions d' expressions verbales. Nous avons décidé d'écarter les expressions à tête nominale ou adjectivale : il y a très peu dans notre base, la plupart sont identifiables à partir d'une liste limitée de valeurs.

Types d'expressions polylexicales à annoter

On annote des expressions polylexicales (qui sont composées d'au moins deux unités lexicales distinctes), composées à partir d'un verbe. Ces unités lexicales sont en relation de dépendance syntaxique (sujet, objet direct, objet indirect complément circonstanciel) et le sens est plus ou moins compositionnel. Plusieurs types d'expressions sont à annoter :
1) expressions idiomatiques : ces expressions ont un sens figuré, non déductible à partir des sens de chaque unité : *jeter un sort, perdre pied* (perdre la confiance en soi), *jeter l'éponge* (abandonner). Ces unités sont caractérisées par un figement syntaxique important (préférence pour un déterminant précis pour le nom : *perdre pied*, mais pas **perdre les pieds* (sauf dans un cas particulier d'un accident, mais dans ce cas précis, il y a un changement de sens) . Par ailleurs, il est impossible d'appliquer un modifieur au nom : **perdre pied gauche* ou au verbe (**perdre rapidement pied*). La passivation est impossible pour les expressions qui ont un verbe comme tête lexicale : **pied a été pris*.
2) collocations : ces expressions manifestent une préférence lexicale forte (*poser une question* mais pas *demander une question), le sens est plus compositionnel. Cette catégorie d'expression manifeste aussi plus de variabilité syntaxique. On accepte des modifieurs adverbiaux (*prendre rapidement des mesures*) ou adjectifs (*prendre une mesure drastique*). Il est possible de passer à la diathèse passive (*les mesures ont été prises par le gouvernement*). Les éléments de la collocation résistent quand on applique des tests de substitution (par exemple on remplace un nom par un synonyme ou un verbe par un synonyme) : le remplacement implique en général un changement de sens. On peut vérifier pour les collocatifs dans Voisins d'Wikipédia http://redac.univ-tlse2.fr/voisinsdewikipedia-5-5-0.1/lemme.jsp?

toDo=chercherLemme&lemme=transmettre, si l'association entre verbe et nom est suffisamment forte.
3) les expressions figées : dans cette catégorie on peut compter les expressions dont le verbe est variable (*être sans espoir pour, avoir droit à*) mais l'objet est totalement fixe ou les expressions figées ayant une valeur pragmatique « pragmatème » (Tutin, 2015) : *ça va s'arranger, le soleil brille, il fait chaud*. Pour ces expressions, on ne peut pas insérer un déterminant, un adjectif ou une relative :(M.Gross, 1993) parle de groupe nominal figé, d'adverbe figé etc. On inclut la préposition si elle est indique un des arguments de l'expression.

Délimitation

Il y a plusieurs cas de figure en cas d'annotation des catégories d'expressions polylexicales étudiées. On peut avoir une expression qui est continue (toutes les unités lexicales sont identifiées comme faisant partie d'une seule expression). Il y a la possibilité d'avoir des expressions discontinues : les unités sont séparées par certains modifieurs, négations. On peut aussi délimiter l'expression comme un seul bloc ou encore identifier toutes les unités, même si elles sont discontinues :

Le lac de Constance donne naissance au Rhin.

Marc pose rapidement une question au professeur d'économie.

Pour délimiter les expressions, nous utilisons les paranthèses [] et un numéro d'expression et une catégorie CAT (IDIOM – expression idiomatique ; COLLOC – collocation ; FIXE – expression figée) et le type de tête (V-pour les verbes ; N- pour les noms ; A-pour les adjectifs)

[expression]_1/CAT/TETE

Exemple 1. L'expression est continue

a) Il **[a gardé la tête froide]_1/IDIOM/V** malgré le danger imminent.
b) À force de parler de lui, il [**a les chevilles qui enflent]_2/IDIOM/V**.
c) C'est lui qui va [**avoir la fève]_3/FIXE/V,** il ne reste qu'une part de galette !
d) Je **[n'ai plus 20 ans]_4/FIXE/V**, je perds vite mon souffle.
e) J'**[ai trouvé un travail]_6/COLLOC/V** incroyable pour le mois de juillet : il faut que je surveille des animaux dans un parc animalier.
f) - Alors, je vais y aller : « Imaginez un pays dont les habitants travaillent trente-cinq heures par semaine, **[ont droit à]_7/FIXE/V** cinq semaines de congés payés par an, prennent des pauses-déjeuner d'une heure et demie, **[ont une espérance de vie]_8/COLLOC/V** des plus longues malgré une tradition culinaire des plus riches.
g) **[Mon sang n'a fait qu'un tour]_9/IDIOM/V**.
h) [**La question que je pose]_10/COLLOC/V.**

Dans le cas des expressions continues, il faut inclure tous les éléments de l'expression, y compris l'auxiliaire du verbe (pour les temps composés). Par contre, les compléments du noms (les relatives, les adjectifs, les groupes prépositionnels) ne seront pas inclus dans l'expression, afin d'avoir plus de chances d'avoir un bon accord inter-annotateur. Dans le cas b), il s'agit d'une expression polylexicale idiomatique qui inclut la relative, pas conséquent on l'annote. Pour les expressions figées (exemple c), en général c'est que le verbe avoir ou être qui va changer, mais qui est considéré comme partie intégrante du verbe. Dans le cas d) on peut avoir aussi une autre valeur pour les nombres (je n'ai plus 18 ans, je n'ai plus 30 ans etc.), le nombre variable est inclus dans

l'annotation. Dans l'exemple e) nous avons inclut aussi l'auxiliaire du verbe pour la collocation (trouver un travail). Dans l'exemple f) on a noté toutes les expressions que nous avons identifiées, y compris la préposition à (pour avoir droit). Dans l'exemple h) on annote l'expression complète.

Exemple 2 . L'expression est discontinue

a) Max **[pose]_12/COLLOC/V** calmement **[une question]_12/COLLOC/V** au fonctionnaire du guichet mais la réponse est très peu rassurante.
b) Je **n'[ai]_13/COLLOC/V** pas encore **[donné mon avis]_13/COLLOC/V** sur la demande de Thierry.
c) **[Le signal]_14/COLLOC/N** fort **[a été transmis]_14/COLLOC/V** à la population.

Dans l'exemple b) il y a la négation qui est séparée de l'annotation de la collocation. L'auxiliaire est annoté ainsi que le déterminant possessif 'mon' (pour d'autres personnes on aurait mis : Ils n'ont pas encore donné **leur** avis).

Problèmes d'expressions trop spécifiques
- vérifier pour les collocatifs dans Voisins d'Wikipédia http://redac.univ-tlse2.fr/voisinsdewikipedia-5-5-0.1/lemme.jsp?toDo=chercherLemme&lemme=transmettre

Problèmes de délimitation
- inclure ou non les prépositions ? Pour certaines expressions, la préposition introduit un constituent obligatoire (être à la charge de, avoir l'air de). Dans ce cas, on doit inclure la préposition. Dans d'autres cas, on peut trouver un verbe équivalent (prendre congé de => quitter) et on annote la préposition.
- ne pas mettre les pronoms réfléchis

être loin de => colloc (comme dans la base)
se mettre à => colloc*
avoir peu

ajouter adopter une loi colloc, contrat conclu colloc
être possible de => non
rester indépedent => non
ne pas manquer de => fixe oui
prendre congé => fixe
servir à dîner =>
être considéré comme => fixe
avoir lieu de => idiomatique
avoir à dire => non
faire honneur à => idiomatique (inclure la préposition si elle est obligatoire)
vider d'un trait => collocation ?
Dégager un parfum => collocation
sentir la mort => collocation
couper le cordon=> idiom
faire l'usage de

tenir dans
émettre le signal
transmettre le signal

rester un hic
assurer la liaison

pas de problème résolu
ne pas être sans rappelle fixe
écarquiller les yeux coloc
passer des tests colloc
donner l'heure collocation
avoir droit à fixe
bras m'en tombent => idiom
faire mine de => idiom
tourner le dos => collocations
faire partie de => collocation
se consacrer à => collocation
trouver son inspiration=> colloc
être fière non
rendre autonome => non
être censé => oui colloc
être ravi => non
retomber sur ses pieds => idioms

Problèmes de catégories

a) différence entre COLLOC et FIXE : les expressions figées sont soit composées avec être ou avoir et suivi d'une expression fixe (être sans espoir, avoir de la peine), leur sens reste compositionnel. Dans cette catégorie d'expressions figées on a aussi « Il fait chaud », « Il fait froid », « c'est terrible » .

b) utilisation d'une expression dans son sens figuré et dans son sens littéral : tourner le dos (littéralement ou figuré). Seulement s'il s'agit du sens figuré, on annote l'expression.

C() verbes supports : plutôt expressions fixes

F-mesure pour MWE

Fspan => on identifie si les unités lexicales composant les expressions qui sont identiques au Gold standard (il faut les retrouver dans l'ordre, après lemmatization)
Fpartiel => on peut compter juste le nombre de tokens en commun (après lemmatization)
Fcat => on identifie la correspondance entre les catégories d'expressions (expression idiomatique, collocation, expression figée)