Guide d'annotation du corpus APPLE (Annotating Possible Puns in Linguistic Expressions)

Présentation du corpus

Notre corpus contient 600 tweets annotés. On y retrouve 2 types d'annotations :

(1) Annotation séquences figées / défigées	on indique pour chaque tweet s'il contient une expression défigée ou non (et si on reconnait l'expression figée sur laquelle le défigement se base).
(2) Annotation du procédé de défigement	on annote les tweets contenant les expressions défigées afin de 1) extraire le défigement et 2) extraire des informations pertinentes pour un futur traitement automatisé.

Nous avons déjà réalisé l'annotation (1), ce qui nous a permi d'identifier **138 tweets contenant une séquence défigée** pour 462 tweets n'en contenant pas. L'annotation (2) se déroule donc sur ces 138 tweets contenant une séquence défigée.

Annotation

Voici les objectifs de notre annotation (2) :

- elle permet de vérifier que ce que nous avons annoté comme défigé est correct (exemple : "le beurre, l'argent du beurre et le cul de la crémière" a parfois été annoté comme défigement de "le beurre et l'argent du beurre", ce qui me semble incorrect) ;
- elle permet d'extraire le défigement du tweet ;
- elle permet d'annoter les modifications occurant entre une séquence figée et une séquence défigée.

Nous avons planifié une annotation au token. Pour chaque token, on annote s'il appartient à une séquence défigée et si oui, on peut préciser s'il s'agit d'une modification d'un mot de la séquence figée d'origine, d'une substitution ou d'une insertion. Nous présentons dans le tableau suivant toutes les annotations disponibles pour un token.

Dernière précision : comme Gaël l'a remarqué, on a parfois une autre expression dans que celle qu'on étudie dans un tweet. Pour ce genre de cas (et pour les cas qu'on ne peut pas traiter avec le schéma d'annotation mis en place), j'ai rajouté une ligne "remarque". Tous les tweets contenant une remarque seront abordés lors de nos réunions d'adjudication, donc n'hésitez pas à faire remonter des incohérences / vos questions dans cette ligne.

Aucun changement (X)	un mot est le même entre la séquence figée d'origine et la séquence défigée.
Modification (M)	un mot du figement a subi une modification formelle et sémantique dans le défigement (exemple : "que" qui devient "qu'" ne pourra pas être considéré comme une modification, une conjugaison différente ou un changement de temps non plus).
Substitution (S)	un mot de la séquence figée est remplacé par un autre mot.
Insertion (I)	un mot qui n'était pas présent dans la séquence figée a été ajouté dans la séquence défigée.
Suppression	on ne l'annote pas, on la constatera en aval en comparant les deux séquences (séquences figée et défigée).
Renversement d'ordre	IDEM, on le constatera en aval en comparant les deux séquences (séquences figée et défigée).
Erreur (E)	Erreur d'orthographe / de grammaire.

Fichier d'annotation

Le fichier CSV prévu pour l'annotation n'est pas terrible, je préfère être honnête avec vous. On va essayer de faire avec et on verra comment on peut l' améliorer / le changer par la suite. En gros le fichier suit la structure suivante :

1	Tweet 1 entier						
2	Seed 1 entière						
3	token 1	token 2	token 3	token 4	token 5	token 6	token 7
4	annot 1	annot 2	annot 3	annot 4	annot 5	annot 6	annot 7
5	Remarque						
6							
7	Tweet 2 entier						
8	Seed 2 entière						
9	token 1	token 2	token 3	token 4	token 5	token 6	
10	annot 1	annot 2	annot 3	annot 4	annot 5	annot 6	
11	Remarque						

Exemples d'annotations

Les exemples sont tirés de 1) notre corpus, 2) notre imagination et 3) la thèse de Lichao Zhu.

Substitution

Dans le cas d'une substitution, on précise quel est le mot substitué au moment de l'annoter. Prenons l'exemple 1. "moins" remplace "plus", donc on annote "S" pour substitution et on précise qu'il substitue "plus" : "S:plus".

1	Qui veut travailler moins pour gagner plus ?								
2	travailler plus pour gagner plus								
3	Qui	veut	travailler	moins	pour	gagner	plus	?	
4			X	**S:plus**	X	X	X		
5									

Exemple 1

Quand on a un mot A qui est modifié de manière à donner un autre mot B mais qu'on reconnait toujours ce mot A dans la séquence défigée, on peut l'annoter "MS", ce qui signifie modification + substitution. Voir l'exemple 2. On a "pote" qui est substitué par "despote", qui est graphiquement et phonétiquement proche.

1	Touche pas à mon despote								
2	Touche pas à mon pote								
3	Touche	pas	à	mon	despote				
4	X	X	X	X	**MS:pote**				
5									

Exemple 2

On fait ainsi la différence entre deux types de substitutions menant à des défigements : celles jouant sur la morphologie et/ou la phonétique d'un mot ("pote" qui devient "despote" dans l'exemple 2, similaires morphologiquement et phonétiquement) et celles jouant sur d'autres procédés (sémantique pour "moins" substituant "plus" dans l'exemple 1).

Question 1: est-ce nécessaire de distinguer ces deux types de substitution et surtout de les annoter ? Je ne pense pas que ça rajoute beaucoup à la complexité de l'annotation et ça ajoute une information que je considère pertinente dans le cas d'étude des défigements.

Question 2 : au final, on ne retrouvera pas de modification (M) seule. Car un mot semble être modifié afin de créer un nouveau mot. Donc la modification apparaît plus comme une indication ajoutée en annotant une subtitution : la substitution est-elle obtenue par rapprochement morphologique / phonétique avec un mot de la séquence figée d'origine ?

Insertion

1	casse-toi pauvre gros con !								
2	casse-toi pauvre con !								
3	Casse	toi	pauvre	gros	con	!			
4	X	X	X	**I**	X				
5									

Exemple 3

Renversement d'ordre

On capture ce changement sans avoir besoin de faire annoter l'ordre des mots par les annotateurs. Dans l'exemple 4, "avec" n'aura pas la même place dans la séquence figée et dans la séquence défigée : il apparait AVANT les mots "la force soit" alors qu'il se place APRES ces mots là dans la séquence figée d'origine.

1	qu'avec toi la force soit.								
2	que la force soit avec toi								
3	qu	avec	toi	la	force	soit			
4	X	X	X	X	X	X			
5									

Exemple 4

On évite ainsi l'intégration d'une étape laborieuse qu'il serait difficile d'annoter. Idem pour la suppression de mots d'une séquence figée à une séquence défigée : on pourra facilement retrouver les mots manquants en comparant les séquences et en s'aidant de nos annotations.

Florilège de cas complexes

Hésitation entre Substitution et Modification : pourquoi pas les deux ? Dans l'exemple 5, "les" doit-il être annoté comme Substitution (S) ou Modification (M) selon vous ? Si on reprend le raisonnement présenté pour l'exemple 2, on pourrait même annoter "les" avec les deux (MS). C'est cette dernière option que je choisirait personnellement, on a ainsi 2 informations d'annotées : le mot a été substitué par un autre mot MAIS ces deux mots sont proches.

1	Rachida garde plus les sceaux								
2	garde des sceaux								
3	Rachida	garde	plus	les	sceaux				
4		X	I	MS:des	X				
5									

Exemple 5

Changement d'ordre des mots de l'énoncé, suppression de mots et une substitution : L'exemple 6 illustre un cas extrême selon moi, avec une nouvelle expression qui nait de l'expression figée d'origine. La moitié de l'énoncé n'est plus présent, le verbe n'est plus le même et les mots ne sont plus dans le bon ordre.

1	Il n'a encore vendu que la peau de l'ours										
2	il ne faut pas vendre la peau de l'ours avant de l'avoir tué										
3	il	n	a	encore	vendu	que	la	peau	de	l	ours
4	X	X	S:faut	I	X	I	X	X	X	X	X
5											

Exemple 6

Question 3 : dans l'exemple 6, le "il" de la séquence figée et le "il" de la séquence défigée ne connotent pas la même chose. L'annoter ? Si oui, comment ?

Concaténation de deux mots en un seul mot : Dans L'exemple 7, on peut préciser qu'il s'agit d'un mot subtituant deux mots : "S:sans+peur". Là encore, annote-t-on "S:sans+peur" ou "MS:sans+peur" ? Après tout, "sans peur" et "sapeur" sont phonétiquement et graphiquement proches.

1	sapeur et sans reproche								
2	sans peur et sans reproche								
3	Sapeur	et	sans	reproche					
4	MS:sans + peur	X	X	X					
5									

Exemple 7

Un mot qui est remplacé par plusieurs mots : Dans l'exemple 8, on a clairement un syntagme ("le peuple français") qui se substitue à un mot ("toi"). Comme plusieurs mots se substituent au mot "toi", on annote tous ces mots de la même manière.

1	que la force soit avec le peuple français								
2	que la force soit avec toi								
3	que	la	force	soit	avec	le	peuple	français	
4	X	X	X	X	X	S:toi	S:toi	S:toi	
5									

Exemple 8

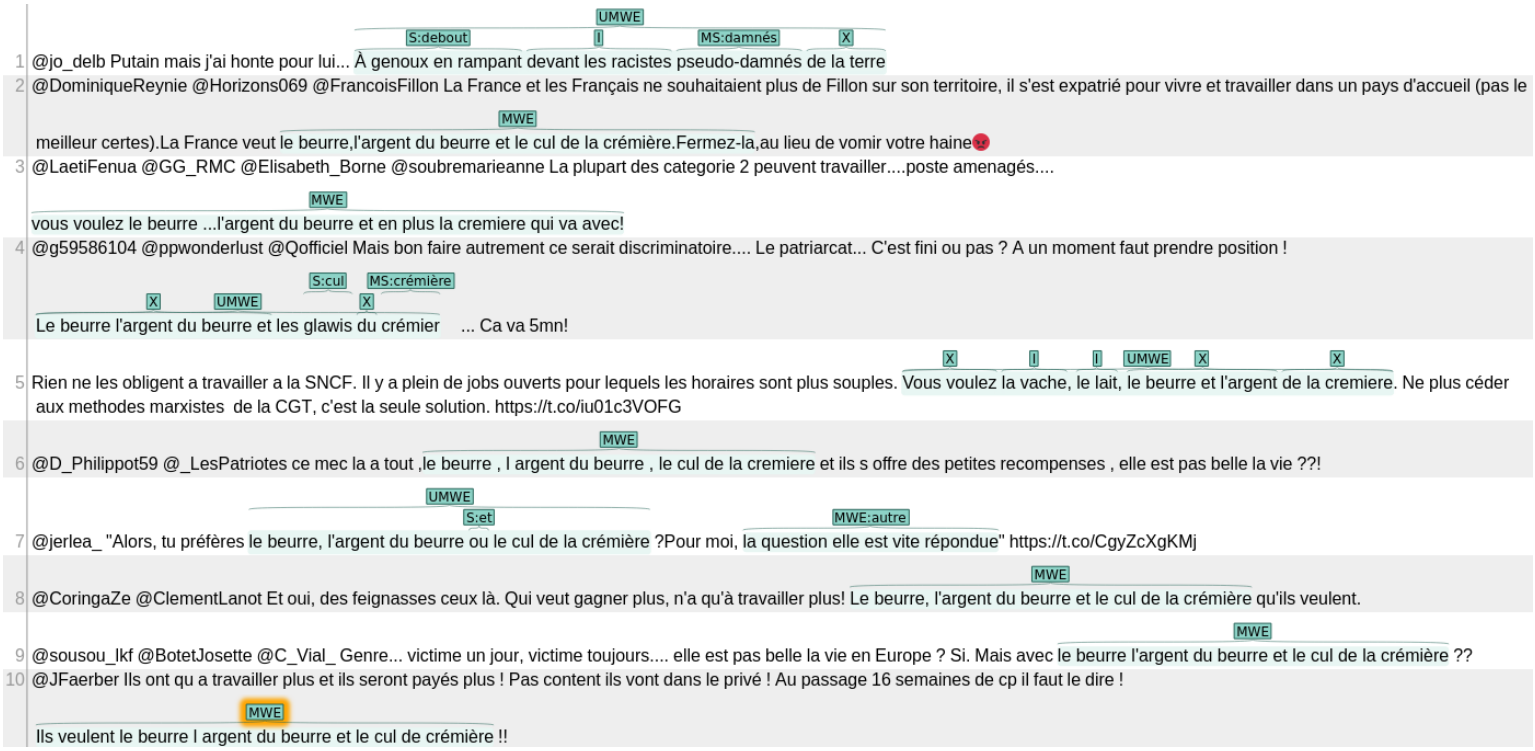
D'autres exemples viendront afin de représenter un maximum de cas. Envoyez-moi des annotations que vous trouvez intéressantes si vous en voyez !

Bonus : autres pistes d'annotation

Lors de mes recherches, je me suis demandé qu'elle serait la meilleure manière d'attaquer cette tâche d'annotation. Comme nous avions réalisé notre première phase d'annotation sur des fichiers CSV, mon premier instinct a été de continuer dans cette voie et de privilégier ce format. Mais est-ce réellement un format adapté à l'annotation que nous souhaitons réaliser ? Voici quelques remarques que j'ai sur ce format :

Choix de l'analyse en tokens	Il nous faut une même unité sur laquelle baser nos annotations. Le choix du caractère me paraît trop compliqué à mettre en place dans un fichier au format CSV. Le choix de la phrase ne me paraît pas adapté 1) au format CSV 2) à une annotation de séquences défigées dans des tweets. Une segmentation en token s'intègre bien aux CSV + permet une annotation assez fine des séquences défigées, même si on peut imaginer des cas où elle ne suffit pas. De toute manière je ne vois pas de solution parfaite pour la segmentation dans notre cas.
Segmentation en tokens	Pour le moment, nous utilisons la segmentation utilisée dans ma méthodologie empirique : retrait de certains caractères problématiques + tokenisation avec SPACY. Je ne suis pas satisfait de la segmentation pour le moment. En segmentant comme je le fait, on introduit un problème : ça contraint l'annotation, on donne aux annotateurs les unités qu'ils doivent annoter sans leur laisser la possibilité de les délimiter eux-mêmes.
Visibilité de l'intégralité du tweet	En proposant le tweet aux annotateurs, on leur laisse l'occasion de repérer dans quelle partie du tweet se situe la séquence défigée à analyser. On a donc un gain de temps, car au lieu de lire le tweet token par token un annotateur peut se diriger directement vers la partie du tweet contenant la séquence défigée.
Visibilité de la seed	En offrant aux annotateurs la seed on s'assure qu'ils ont toujours sous les yeux la séquence figée à laquelle comparer la séquence défigée afin de parfaire le analyse.

J'ai regardé du côté des outils d'annotation et j'ai choisi d'en tester sommairement un afin de voir si ce genre d'outil pourrait être une bonne alternative. Mon choix s'est porté sur INCEpTION (https://inception-project.github.io/) qui est très facile à installer + à lancer. Voici un exemple de schéma d'annotation effectué en quelques minutes, suivit de mes remarques.



Plus de contrôle sur les unités à annoter	Les annotateurs délimitent eux-mêmes les unités qui bénéficient d'une annotation, ce qui peut être un plus. Cependant, on perd l'alignement entre les annotateurs : ils n'annoteront plus systématiquement les mêmes unités vu qu'ils les délimitent eux-même.
Visibilité de l'intégralité du tweet ++	C'est mieux qu'avec le CSV puisque le tweet qu'on voit en entier est aussi la chaîne dans laquelle on va faire notre annotation
Pas de visibilité de la seed	On peut peut-être changer ça.
Temps d'annotation + long	Lors de mon test, j'ai pris plus de temps à annoter 10 tweets que 20 tweets avec le CSV. Il faut cliquer, sélectionner la séquence à annoter et effectuer l'annotation dans le menu contextuel.

Je n'ai pas encore tout vu sur cet outil, mais il existe certainement d'autres fonctionnalités qui peuvent être bénéfiques pour nous. Cependant, perdre l'alignement qu'offre une segmentation préalable me paraît embêtant. Je vais continuer mon exploration sur cet outil pour voir...