Classification automatique de questions spontanées vs. préparées dans des transcriptions de l'oral

L'article étudié porte sur les différences entre les questions spontanées et préparées, le but étant de créer un modèle permettant la classification automatique de celles-ci.

Les énoncés interrogatifs utilisés dans cet article sont extraits de transcriptions d'entretiens guidés des corpus ESLO2 (enquête sociolinguistique à Orléans 2) et ACSYNT. Ce sont des corpus oraux (enregistrement de la parole) français contemporains.

Nous avons d'abord tenté de constituer un corpus similaire en extrayant des questions de podcasts, talk show et autres contenus audio transcrits avec des parties d'interview. Comme il était difficile de déterminer quelles étaient les parties préparées et spontanées et que les formats de transcription varient trop d'un contenu à l'autre, nous avons pris un autre corpus constitué d'interviews de parents sur l'aide qu'ils apportent à leur enfant dans son apprentissage de la lecture. L'avantage de ce corpus est que la trame nous est donnée, ce qui facilite le repérage des questions préparées. Il est composé de 84 interviews de parents. Nous avons d'abord converti les documents pdf de chaque interview en documents txt qui sont plus faciles à manipuler. Ces documents ont ensuite été fusionnés pour obtenir un seul document txt.

Dans l'article les énoncés à classifier sont relevés en se servant du point d'interrogation comme marqueur. Ils sont divisés en trois catégories: "spontané", "préparé" et "non-question". Les 5 tours de paroles précédant et suivant chaque question sont extraits et servent de contexte pour aider à annoter la question. Le corpus final est composé de 1298 énoncés annotés pour ESLO et 588 pour ACSYNT.
Nous avons extrait les questions de façon similaire, en nous servant du point d'interrogation comme marqueur l'énoncé interrogatif, mais nous n'avons pas relevé plusieurs tours de parole. Le corpus étant beaucoup plus petit et plus facilement consultable pour regarder le contexte d'une question en cas de doute, nous avons à chaque fois pris le texte allant du speaker (la personne qui parle) au point d'interrogation. Notre corpus final de questions annotées est constitué de 200 échantillons.

Une question est un acte de dialogue consistant en une recherche d'information. Il existe donc des énoncés se terminant par un point d'interrogation qui ne sont pas des questions, ceux-ci sont annotés "non-question". Les énoncés se terminant par un point d'interrogation qui consistent en une recherche d'information, lorsqu'ils sont compréhensibles (ceux qui ne le sont pas ne sont pas retenus, marqués comme "poubelle", car non annotables) sont alors annotés "spontané" ou "préparé".
La typologie de l'annotation est donc la suivante:

Spontané: Les questions spontanées sont caractéristiques de l'oralité, elles sont construites en direct au cours de la conversation. Parmi les différents marqueurs des questions spontanées, on retrouve les disfluences (hésitations, répétitions de parties de mots, insertion de mots non significatifs à la conversation) une influence du contexte immédiat (la question spontanée peut par exemple servir à obtenir des précisions ou des informations supplémentaires)

Préparé: Les questions préparées sont en général planifiées à l'avance, en amont de la production du discours. Par exemple, dans un entretien sociologique, les questions préparées sont écrites sur une trame et auront donc souvent une syntaxe et une formulation plus claires que les questions spontanées. Pour ce qui est de leurs caractéristiques récurrentes, elles présentent en général moins de disfluences, une répétition d'entités nommées, un mot interrogatif en position initiale. Elles peuvent aussi être annoncées ("une dernière question")
Non-question: inclut les énoncés se terminant par un point d'interrogation qui ne relèvent pas de la recherche d'information. Les non-questions peuvent être par exemple des demandes de clarification, de répétition, d'action, elles servent parfois au bon déroulement de la conversation, par exemple pour vérifier que l'interlocuteur écoute et comprend ce qui lui est dit. Elles peuvent aussi être injonctives (demande d'action) ou servir à contrôler le dialogue.

Dans l'article l'annotation manuelle par deux annotateurs experts (Iris Eshkol-Taravella, Angèle Barbedette) de 200 questions obtient un accord inter annotateur de 0.75.

Pour notre travail nous avons pré-annoté ensemble les questions préparées à l'aide de la trame, en nous servant aussi de critères comme l'annonce de question. Nous avons ensuite chacune de notre côté annoté les questions restantes comme "non question" ou "question spontanée". Nous avons parfois rencontré des difficultés pour différencier certaines non questions d'obligation sociale et des questions spontanées sur les enfants.
Nous obtenons un accord inter annotateur de 0,625.

Certains critères utilisés pour repérer les questions spontanées en français nous ont été utiles, par exemple, dans nos échantillons, nous retrouvons beaucoup de demandes de clarification (exemple: "is that a reward or the other way?"). Nous avons aussi retrouvé de nombreuses anaphores "HV1: Um, so **that** is one of the ways you motivate her?".
D'autres critères n'ont presque pas servi pour notre corpus, même s'ils pourraient être utiles dans d'autres corpus en anglais. Par exemple, nous n'avons eu qu'une fois un mot interrogatif en position finale ("like what?") dans nos échantillons, bien que ce soit une configuration possible en anglais. Même en regardant d'autres questions extraites en dehors de nos échantillons, à part quelques questions uniquement composées d'un mot interrogatif ("what?" "why?"), les questions dans ces interviews ne se terminent presque jamais par des mots interrogatifs. Cela peut-être dû au fait que le mot interrogatif à la fin de la question en anglais est plutôt familier (exemples "and you did this how?").

Autres observations: le mot "so" était présent dans presque une question sur 6, principalement dans les non questions (majoritairement questions d'obligation sociale) et les questions préparées. Dans les questions préparées il semble servir à annoncer/faire une transition vers un nouveau thème, il est souvent placé après un autre mot (exemples: "Nice. **So** what are some things you do to help Child N become a good reader?" "Okay. **So** what are some questions you ask when you talk about books that Child N's read?") tandis que dans les non questions d'obligation sociale, le "so" peut-être vu comme une façon de rendre la conversation plus informelle et établir un ton amical (exemple: "So you finished reading, right?").