Schémas Winograd en français: une étude statistique et comportementale

Pascal Amsili Olga Seminck
Laboratoire de Linguistique Formelle
Université Paris Diderot
amsili@linguist.univ-paris-diderot.fr,
olga.seminck@cri-paris.org

RÉSUMÉ

Nous présentons dans cet article une collection de schémas Winograd en français, adaptée de la liste proposée par Levesque *et al.* (2012) pour l'anglais. Les schémas Winograd sont des problèmes de résolution d'anaphore conçus pour être IA-complets. Nous montrons que notre collection vérifie deux propriétés cruciales : elle est robuste vis-à-vis de méthodes statistiques simples ("Google-proof"), tout en étant largement dépourvue d'ambiguïté pour les sujets humains que nous avons testés.

ABSTRACT

Winograd schemas in French : a statistical and behavioral study

We present in this paper a collection of Winograd schemas in French, adapted from the English collection proposed by Levesque *et al.* (2012). Winograd schemas constitute anaphora resolution problems meant to be AI-complete. We show that our collection has two crucial properties : it is robust regarding simple statistical techniques of resolution ("Google-proof") and basically non ambiguous for the human participants that were tested.

MOTS-CLÉS : Résolution d'anaphores, schémas Winograd, information mutuelle, test de Turing.

KEYWORDS: Anaphora resolution, Winograd Schemas, Mutual Information, Turing test.

1 Les schémas Winograd

Les schémas Winograd [1], proposés par Levesque *et al.* (2012), se présentent comme des cas particuliers du problème de résolution anaphorique, qui ne peuvent être résolus qu'en faisant appel à un raisonnement sur des connaissances du monde. Un schéma Winograd est constitué d'une paire de discours identiques à un mot (ou une expression) près, et qui comprennent une expression anaphorique à résoudre, dont l'antécédent change d'une version à l'autre. Ainsi, dans (1), la réponse naturelle pour la question formée avec le mot faible (on appellera ce mot *special*) est *Nicolas* (R0), alors que la question formée avec le mot lourd (appelé *alternate*) appelle plutôt la réponse *son fils* (R1).

(1) Nicolas n'a pas pu soulever son fils car **il** était trop ⟨faible/lourd⟩.
Qui était trop ⟨faible/lourd⟩ ? R0 : Nicolas
R1 : son fils

1. Ce nom a été proposé en référence à une paire d'exemples proposés par Winograd (1972) pour illustrer la difficulté du problème de la compréhension du langage naturel.

L'intérêt de disposer d'une telle paire de discours quasi identiques est de garantir que la structure linguistique n'empêche aucun syntagme nominal d'être un antécédent possible pour le pronom. Un schéma Winograd regroupe donc deux questions liées que nous appellerons *items* dans la suite.

Parce qu'ils nécessitent des capacités de raisonnement et d'intégration de connaissances encyclopédiques, les schémas Winograd ont été proposés comme un nouveau test d'intelligence artificielle dans l'esprit du fameux test de Turing (1950) (où un système automatique doit convaincre un juge humain qu'il est humain en discutant avec lui). En effet, les schémas Winograd évitent deux écueils du test de Turing : il n'est pas nécessaire pour les résoudre de faire semblant d'être humain (dans le test de Turing, la machine doit prétendre avoir tous les traits d'un être humain, avoir une enfance, une couleur préférée...), et de plus les schémas ne peuvent faire l'objet du genre de stratégie d'évitement qu'une machine pourrait mettre en œuvre (changement de sujet de la conversation, réponse par une autre question, etc.).

Suite à la publication des premiers schémas Winograd (Davis *et al.*, 2015), a été organisé en 2016 le premier *Winograd Schema Challenge* (Morgenstern *et al.*, 2016). Ce concours comportait au premier tour une série de problèmes de résolution de pronoms ambigus inspirés des schémas Winograd, mais qui pouvaient comporter plus de deux antécédents possibles, et qui n'étaient pas aussi contrôlés que les items issus de schémas Winograd (2).

(2) The storekeepers stayed in town to run their stores and lived in the rooms behind **them**.
A : storekeepers
B : stores
C : rooms

Selon les règles du concours, un second tour ne comportant que des vrais schémas Winograd était prévu au cas où le taux de réussite au premier tour dépassait les 90%, mais ça n'a pas été le cas en 2016. Parmi les 5 systèmes soumis au concours, le meilleur (Liu *et al.*, 2016) a obtenu un score de 58% d'exactitude, ce qui montre la difficulté de la tâche [2]. Par la suite, les mêmes auteurs ont élaboré une nouvelle version de leur système significativement meilleure, qui a obtenu le taux de 66,7%. Plusieurs autres travaux portant sur les schémas Winograd ont été publiés en dehors du concours, mais ils ne portent que sur un sous-ensemble des schémas (entre 5 et 25%) choisis pour le rôle qu'y jouent divers phénomènes comme la causalité, la pertinence, ou les relations de discours (Bailey *et al.*, 2015; Schüller, 2014; Sharma *et al.*, 2015, *i.a.*), et leur traitement implique une phase manuelle.

Deux propriétés des schémas Winograd nous intéressent : d'une part ils doivent nécessiter de l'intelligence artificielle, ce qui veut dire en particulier qu'ils ne doivent pas pouvoir être résolus avec des méthodes statistiques simples (on parle de *Google-proofness*), et d'autre part ils doivent être faciles pour les humains. Dans cet article, nous présentons une collection de schémas en français, adaptée de la collection originale de Davis *et al.* (2015), et rapportons les résultats de deux études : la première établit que nos schémas sont résistants à un test statistique basé sur l'information mutuelle ; la seconde est une étude comportementale qui permet de montrer que les humains n'ont pas de difficulté pour traiter correctement les items de notre collection. Notre test statistique simple ne parvient pas à dépasser 55% de taux de réussite alors que les humains obtiennent un taux de 93%.

2. Le nombre d'antécédents possibles étant de 2,33 en moyenne sur les 60 problèmes, le taux de réussite théorique pour une réponse au hasard était de 42,8%.

2 Collection de schémas en français

Notre collection a été produite en prenant comme point de départ les schémas de la collection anglaise de 144 schémas [3] (Davis *et al.*, 2015). Comme ces schémas ont été traduit en japonais et partiellement en chinois, nous avons décidé de les prendre comme point de départ pour permettre une comparaison multi-langue. Le principe a été de tenter une traduction directe des schémas, mais dans la plupart des cas il a fallu procéder à une adaptation, ne serait-ce qu'à cause des propriétés de genre et de nombre : les deux antécédents doivent être compatibles avec le pronom. Nous illustrons dans cette section quelques-uns des problèmes qui se sont présentés.

Dans le cas d'un item comme (3a), la traduction directe n'est pas possible, puisque le singulier *hair* se traduit en un pluriel (*cheveux*). Pour résoudre le problème, nous avons cherché un autre mot qui pouvait correspondre en nombre, et pour cet exemple particulier nous avons remplacé *cheveux* par *savon* (3b). Cet exemple illustre aussi le fait que nous avons cherché à rendre l'exemple aussi naturel que possible, au prix d'une distance plus grande entre l'original anglais et notre version.

(3) a. The drain is clogged with hair. **It** has to be ⟨cleaned/removed⟩.
b. Il y a du savon dans le siphon de douche. Il faut **le** ⟨retirer/nettoyer⟩.

Divers schémas présentaient des problèmes d'ordre lexical, comme par exemple l'item (4a) : la traduction directe de *indiscreet* donne *indiscrète*, qui se trouve avoir (au moins) deux sens distincts en français : en plus de désigner une personne qui révèle des secrets (le sens du mot *indiscreet*), le mot désigne une personne qui cherche à découvrir des secrets (le sens du mot *nosy*). Le schéma a donc dû être entièrement adapté :

(4) a. Susan knows all about Ann's personal problems because **she** is ⟨nosy/indiscreet⟩.
b. Sylvie est au courant de tous les problèmes personnels de Marie car **elle** est ⟨curieuse/bavarde⟩.

On peut encore évoquer le cas de l'item (5a) qui n'a pas pu être adapté, à cause de la préférence en français pour une structure infinitive pour exprimer une subordonnée finale portant sur le sujet de la principale : la version (5b) est fortement dégradée.

(5) a. Mary tucked her daughter Anne into bed, so that **she** could ⟨sleep/work⟩.
b. #Mary$_i$ a mis sa fille au lit (pour/afin/de sorte) qu'**elle**$_i$ dorme.

Le travail de traduction a été réalisé par plusieurs stagiaires, séparément puis en concertation, avant une validation par les auteurs. La traduction la plus naturelle (et si possible pas trop longue) a toujours été privilégiée, et les items sur lesquels il n'y avait pas d'accord entre les traducteurs ont été exclus. La collection finale comprend 107 schémas (214 items) et peut être librement téléchargée sur le site suivant : `http://www.llf.cnrs.fr/winograd-fr`. Chaque schéma comprend un lien vers le schéma original en anglais.

3. Au moment de nos expériences, il y avait 144 schémas, mais aujourd'hui la collection compte 146 schémas.

3 Test de robustesse statistique

L'enjeu des schémas Winograd est donc qu'il doit être nécessaire pour les résoudre de raisonner sur des connaissances du monde. Cela implique qu'il devrait être impossible d'avoir de bons résultats en appliquant de simples tests statistiques (Levesque *et al.* (2012) parlent de la propriété de *Google-proofness*). Ainsi, par exemple, on peut penser qu'un item comme (6) n'est pas *Google-proof*, parce qu'il suffirait d'exploiter des statistiques de co-occurrence en corpus (ou des requêtes à un moteur de recherche) pour déduire qu'il vaut mieux associer *vite* avec *bolide* qu'avec *bus scolaire*.

(6) Le bolide a dépassé le bus scolaire, parce qu'**il** roulait très vite.

Même si notre collection a été adaptée de celle de Davis *et al.* (2015) où les items les plus douteux ont été testés pour la *Google-proofness* (d'après la documentation fournie avec la collection, au moyen de comptages de nombre de *hits* donnés par Google), notre projet était de mener ce test pour la collection française dans son ensemble, de façon rigoureuse et systématique.

Pour cela, nous avons défini un test basé sur l'information mutuelle (Shannon & Weaver, 1949). L'information mutuelle permet de mesurer la dépendance entre deux variables aléatoires, et cette mesure a déjà été utilisée pour mesurer l'association entre deux mots : si les mots x et y sont mutuellement dépendants, la probabilité de leur cooccurrence $P(x, y)$ sera plus élevée que la probabilité d'observer les deux mots ensemble par hasard : $MI(x, y)$ sera positive (Church & Hanks, 1990) :

$$MI(x, y) = log_2 \left(\frac{P(x, y)}{P(x)P(y)} \right) \quad (1)$$

Pour tester la robustesse de nos schémas, nous avons cherché une méthode permettant de choisir entre les deux réponses possibles pour chaque item. Par exemple, pour l'item (7a), le choix doit se faire entre les deux réponses possibles (7b). Nous mesurons l'information mutuelle entre les mots importants de ces deux réponses, ce qui donne (7c).

(7) a. La sculpture est tombée de l'étagère car **elle** était trop encombrée.
b. (R0, alt) la sculpture était trop encombrée
(R1, alt) l'étagère était trop encombrée
c. MI(sculpture, encombrer) = 4,23
MI(étagère, encombrer) = 10,01

La façon la plus simple d'interpréter ces scores est de choisir la réponse qui maximise le score. Pour l'item (7a), cela donne la bonne réponse (*étagère*). Cependant, cette stratégie peut aussi échouer, par exemple sur l'autre item associé au même schéma :

(8) a. La sculpture est tombée de l'étagère car **elle** était trop lourde.
b. MI(sculpture, lourd) = 2,41
MI(étagère, lourd) = 4,03

Dans ce cas, la réponse correcte est *la sculpture*, mais ce n'est pas cette réponse qui a le plus fort score. On peut penser que les scores d'information mutuelle sont trop bas, ou plus simplement que c'est l'écart entre les deux scores qui est trop petit pour être significatif. Cela nous a conduit à introduire un seuil d'écart entre les scores, que nous faisons varier pour en étudier l'impact.

Il est important de préciser que notre méthode exige que les schémas aient une certaine forme : le score MI peut être calculé quand les deux réponses possibles comprennent les deux antécédents

possibles et le mot *special* (ou l'*alternate*). La collection comprend des schémas, comme (9a), qui n'ont pas cette propriété : la paire de réponses possibles est la même pour les deux items du schéma.

(9) a. Anna a ⟨mieux/moins bien⟩ réussi l'examen que son amie Lucy car **elle** avait beaucoup révisé.

b. (R0, ⟨spe/alt⟩) Anna a beaucoup révisé
(R1, ⟨spe/alt⟩) Lucy a beaucoup révisé

Nous avons exclu de notre analyse les 30 schémas de ce type, ainsi que 2 autres schémas où le jeu de réponses est différent selon l'item associé.

Il convient de noter de plus que dans une grande partie des schémas on trouve des noms propres comme antécédents possibles, ce qui donne par construction des items *Google-proof*, puisque dans le cas général les co-occurrences de noms propres avec des noms communs sont aléatoires. Néanmoins, quand il nous était possible de mesurer les fréquences pour ces items, nous les avons inclus dans nos scores. Au total, nous avons mesuré l'information mutuelle pour 90 schémas (180 items).

Pour estimer MI nous avons utilisé les mesures de fréquence (sans lissage) du corpus FrWaC (Baroni *et al.* (2009), 1,6 millards de mots du domaine .fr d'Internet). Sauf quand la différence portait sur la flexion, nous avons utilisé les lemmes pour les mesures, et nous avons choisi les têtes dans le cas d'expressions poly-lexicales. Nous avons choisi d'utiliser un corpus à taille fixe, au lieu des comptages Google, car ces derniers ne sont pas stables (Lapata & Keller, 2005). La Figure 1 donne le taux de réussite de notre méthode pour les différents seuils. Sur les 180 items de départ, 49 n'ont pas pu recevoir un score car une des mesures de fréquence/co-occurrence était nulle.

Seuil	Nb items	Exactitude	Couverture
Aucun	131	0,55	0,40
Δ 0,5	95	0,59	0,31
Δ 1,0	73	0,62	0,25
Δ 1,5	59	0,64	0,21
Δ 2,0	38	0,68	0,14
Δ 2,5	30	0,70	0,12
Δ 3,0	25	0,68	0,09
Δ 3,5	18	0,67	0,07
Δ 4,0	15	0,60	0,05

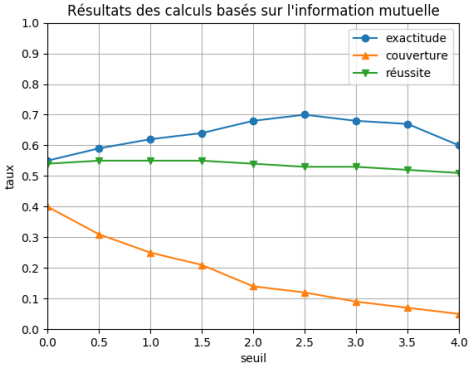

FIGURE 1 – Résultats des calculs d'information mutuelle. L'*exactitude* est le nombre de réponses correctes pour les items auxquels la méthode s'applique, alors que la *couverture* correspond au nombre de réponses correctes divisé par le nombre total d'items (180). Par *réussite* on désigne le taux de réussite théorique qui serait obtenu par une stratégie consistant à utiliser les scores d'information mutuelle quand la différence est supérieure au seuil, et à répondre au hasard dans les autres cas.

Rappelons qu'une réponse au hasard à l'ensemble des question donnerait un taux de réussite autour de 50%. Par conséquent, quand aucun seuil n'est appliqué, le taux obtenu (55%) n'est pas vraiment différent du hasard. Quand le seuil de Δ 2,5 est appliqué, le taux de réussite monte à 70%, mais le nombre d'items auxquels la méthode s'applique est extrêmement petit : moins de 15%. La courbe *réussite* montre que quel que soit le seuil choisi, on ne parvient pas à dépasser 55% de réussite globale.

On peut donc conclure que notre collection est *Google-proof*, au moins dans le sens où nous avons interprété ce terme.

4 Résolution pour les sujets humains

Pour que les schémas jouent leur rôle de test d'IA, ils ne doivent pas poser de difficulté pour les humains. Pour vérifier que notre collection était correcte de ce point de vue, nous avons lancé une expérience en ligne (sur *Ibex Farm*) où nous avons présenté les items aux participants, en leur demandant de répondre à la question. Nous avons garanti que chaque participant était exposé à un seul des deux items de chaque schéma, et contrebalancé l'ordre de présentation. Au total, nous avons testé 203 items [4] sur 22 participants. Nous avons exclu de l'analyse les réponses associées à un temps de réaction inférieur à $1''$ ou supérieur à $60''$.

La réussite moyenne de nos participants est de 93,6%, ce qui est très proche du chiffre de 92% établi pour les schémas anglais (Bender, 2015). Ce taux moyen cache une certaine disparité que l'on peut mettre en évidence en étudiant le taux moyen par item. La Figure 2 donne le nombre d'items en fonction de leur taux moyen de réussite. Si beaucoup d'items sont très faciles (145 items ont un taux élevé de 80 à 100%), il reste 19 items dont le taux est inférieur à 80%, et nous avons étudié individuellement ces items problématiques. Nous proposons de distinguer trois classes de schémas difficiles.

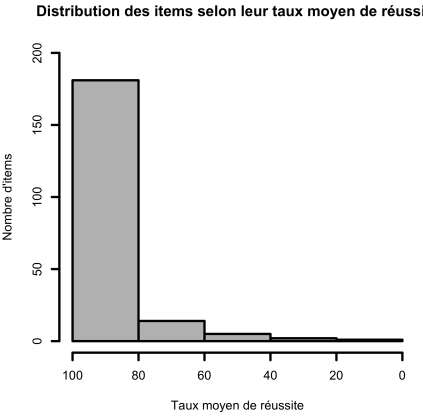

FIGURE 2 – Le taux de réussite moyen pour les participants humains est de 93,6% sur l'ensemble des items testés. L'analyse par item montre que la plupart des items ont un taux très élevé, et sont donc très faciles.

La première classe est constituée d'items ambigus. Ce genre de schéma ambigu est une conséquence du caractère artificiel de la construction des schémas Winograd : en cherchant un schéma, on perd de vue que dans une situation où seule une version est présentée, l'item peut être ambigu. Le schéma (10) illutre ce cas (les deux items sont mal réussis : *special* : 50%, *alternate* : 18%).
L'ambiguïté concerne 7 items sur les 19 étudiés.

(10) Pierre et Marc sont poursuivis pour diffamation. Pierre a écrit dans leur livre plusieurs faux témoignages que Marc a colportés. **Il** aurait dû être plus ⟨prudent/honnête⟩.

Une deuxième classe d'items à problème comprend des items où un connecteur discursif joue un rôle important. Dans le schéma (11) le connecteur *alors que* introduit une opposition. Or, on constate que cette opposition n'est pas toujours interprétée : la version *special* du schéma est réussie à 70% et l'*alternate* à 75%. Ce problème de connecteur concerne 6 items.

4. Le nombre d'items est différent de celui la collection finale, car cette expérience a été conduite avant la toute dernière version de notre collection.

(11) Les pompiers sont arrivés ⟨avant/après⟩ les policiers alors qu'**ils** venaient de plus loin.

Une dernière catégorie concerne des items qui ont l'air d'être « trop construits », pas naturels. Par exemple le schéma (12) pour lequel le *special* a un taux de réussite à 50% et l'*alternate* 67%. Nous avons classé au total 6 items dans cette dernière catégorie.

(12) Pierre jouait aux cartes avec Adam qui menait au score. Si la chance d'Adam n'avait pas tourné **il** aurait ⟨perdu/gagné⟩.

En excluant ces items difficiles de notre collection, nous obtiendrions certainement un taux de réussite supérieur à 93,6%, mais ce score en tant que tel, compte tenu des erreurs inhérentes à ce genre d'étude comportementale, peut déjà être considéré comme très proche d'un score parfait.

5 Conclusion

Nous étudions dans cet article la collection de schémas Winograd en français que nous avons adaptée de la collection anglaise (Levesque *et al.*, 2012; Davis *et al.*, 2015). Cette étude montre que nos schémas sont *Google-proof* : un test statistique simple ne permet pas de dépasser un taux de réussite de 55%, à comparer avec le taux de 93,6% obtenu par les participants à une étude comportementale sur les mêmes données.

À partir de ces résultats nous pouvons envisager de produire une collection encore plus *Google-proof* et encore plus naturelle pour les humains, en ôtant les items repérés comme à la limite dans les deux cas et en ajoutant de nouveaux schémas, sans faire appel à la traduction. Nous envisageons aussi de profiter de cette occasion pour préparer des items en français pour le prochain *Winograd Schema Challenge*.

Remerciements

Nous remercions nos stagiaires du cursus Linguistique Informatique à l'Université Paris Diderot, Sarah Ghumundee, Biljana Knežević, et Nicolas Bénichou, pour leur aide dans la préparation des items et le calcul des scores d'information mutuelle. Nous remercions également les étudiants de la Licence Frontières du Vivant au Centre de Recherches Interdisciplinaires, Ryan Hunt, Dara Nguyen et Hugo Taquet pour leur étude comportementale préliminaire. Nous remercions aussi les relecteurs de TALN 2017 qui nous ont aidé à améliorer cet article, ainsi que les relecteurs et les participants de l'atelier CORBON associé à EACL 2017 (Valence) où une version préliminaire de ce travail a été présentée (Amsili & Seminck, 2017). Ce projet est soutenu en partie par l'École Doctorale Frontières du Vivant — Programme Bettencourt.

Références

AMSILI P. & SEMINCK O. (2017). A Google-proof collection of French Winograd Schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, p. 24–29, Valencia : EACL Association for Computational Linguistics.

BAILEY D., HARRISON A., LIERLER Y., LIFSCHITZ V. & MICHAEL J. (2015). The winograd schema challenge and reasoning about correlation. In *In Working Notes of the Symposium on Logical Formalizations of Commonsense Reasoning*.

BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, **43**(3), 209–226.

BENDER D. (2015). Establishing a human baseline for the winograd schema challenge. In *MAICS*, p. 39–45.

CHURCH K. W. & HANKS P. (1990). Word association norms mutual information, and lexicography. *Computational Linguistics, Volume 16, Number 1, March 1990*.

DAVIS E., MORGENSTERN L. & ORTIZ C. (2015). A collection of winograd schemas. `http://www.cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html` – Web page collecting 144 Winograd pairs, with comments and references.

LAPATA M. & KELLER F. (2005). Web-based models for natural language processing. *ACM Transactions on Speech and Language Processing (TSLP)*, **2**(1), 3.

LEVESQUE H., DAVIS E. & MORGENSTERN L. (2012). The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

LIU Q., JIANG H., LING Z.-H., ZHU X., WEI S. & HU Y. (2016). Combing context and commonsense knowledge through neural networks for solving winograd schema problems. *arXiv preprint arXiv :1611.04146*.

MORGENSTERN L., DAVIS E. & ORTIZ JR. C. L. (2016). Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, **37**(1), 50–54.

SCHÜLLER P. (2014). Tackling winograd schemas by formalizing relevance theory in knowledge graphs. In *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

SHANNON C. E. & WEAVER W. (1949). The mathematical theory of information.

SHARMA A., VO N. H., ADITYA S. & BARAL C. (2015). Towards addressing the winograd schema challenge-building and using a semantic parser and a knowledge hunting module. In *Proceedings of Twenty-Fourth International Joint Conference on Artificial Intelligence. AAAI*.

TURING A. M. (1950). Computing machinery and intelligence. *Mind*, **59**(236), 433–460.

WINOGRAD T. (1972). Understanding natural language. *Cognitive psychology*, **3**(1), 1–191.