La phylogénie des langues au service de l'analyse automatique

Mathieu Dehouck, Pascal Denis

La phylogénie des langues au service de l'analyse automatique

Récemment recruté comme chercheur CNRS au sein du laboratoire Langues, Textes, Traitements informatiques, Cognition (LATTICE, UMR8094, CNRS / Université Sorbonne Nouvelle / ENS Paris), Mathieu Dehouck s'intéresse à l'analyse syntaxique automatique des langues, en particulier dans une perspective multilingue. Pascal Denis est chercheur et responsable adjoint de l'équipe Inria MAGNET au sein du Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL, UMR9189, CNRS / Université de Lille / Centrale Lille). Ses recherches portent notamment sur l'apprentissage automatique pour le traitement automatique des langues, et ses applications pour l'analyse de la phrase et du discours.

Apprendre de l'histoire des langues

Les spécialistes en linguistique historique et en phylogénie essayent de reconstruire l'histoire des langues et de leur évolution en recoupant des données venant de plusieurs langues plus ou moins proches. Notre travail suit le chemin inverse, à savoir : comment l'histoire des langues, représentée sous forme d'un arbre phylogénique, peut-elle nous informer sur les langues elles-mêmes ?

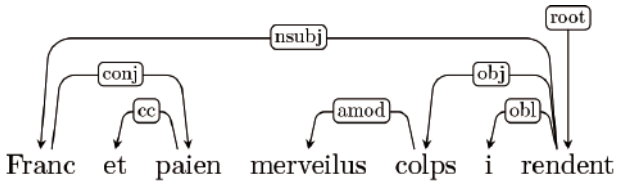

Figure 1 - Un exemple de phrase annotée avec un arbre syntaxique en dépendances. Cette phrase est extraite du corpus Ancien Français SRCMF disponible dans le projet *Universal Dependencies* (Prévost S. and Stein A. (eds.) 2013, *Syntactic Reference Corpus of Medieval French* (SRCMF), ENS de Lyon, ILR Stuttgart.

C'est une question cruciale pour le traitement automatique des langues (TAL), notamment dans la perspective de fournir des outils d'analyse pour le plus grand nombre de langues possibles, et de tenter de résorber la fracture numérique entre les populations. En effet, les quantités de données annotées disponibles pour l'apprentissage automatique de modèles d'analyse sont très inégales. Ainsi, il existe des disparités entre différentes familles de langues et, dans une même famille, certaines peuvent disposer des corpus annotés de plusieurs dizaines de milliers de phrases alors que d'autres en ont beaucoup moins, voire pas du tout.

Dans le cas précis de l'analyse syntaxique en dépendances (Figure 1), sur laquelle portent nos travaux récents, la ressource de référence est le projet *Universal Dependencies*[1], dont la version la plus récente[2] contient 184 corpus pour un total de 104 langues. Les annotations pour ces langues sont réparties de manière très déséquilibrée, avec une majorité de langues indo-européennes et, pour le moment, aucune langue native (d'avant la colonisation européenne) d'Amérique du nord ou du Pacifique sud, par exemple. En outre, même les langues disposant de corpus annotés présentent de grandes disparités de traitement. L'albanais, dispose ainsi d'un corpus de soixante phrases seulement pour un peu moins de 1 000 mots, alors que le tchèque compte cinq corpus totalisant plus de 127 000 phrases et 2, 2 millions de mots. La conséquence de ce déséquilibre est que, à l'heure actuelle, seule une poignée de langues, les mieux dotées en corpus annotées, disposent d'analyseurs syntaxiques suffisamment précis pour donner lieu à des applications réelles.

La typologie sans la typologie

Bien que la conception d'analyseurs syntaxiques procèdent le plus généralement de façon indépendante pour chaque langue, il existe néanmoins déjà quelques approches, dites par transfert, pour créer des modèles pour des langues peu ou non dotées à partir des données de langues bien dotées plus ou moins proches. En revanche, très peu d'approches exploitent l'histoire des langues comme source d'information.

À titre d'exemples, Naseem et al.[3] et Aufrant et al.[4] se sont attaqués au problème de l'apprentissage automatique d'analyseurs syntaxiques en utilisant de l'information typologique (ordre des mots, existence d'articles, existence d'un verbe copule) pour guider le partage de l'information entre les différentes parties de leurs modèles. Ces approches souffrent d'une limitation importante, à savoir : la nécessité d'avoir accès à des informations typologiques fiables sur les langues concernées, qui sont rarement disponibles pour les langues peu dotées.

Pourtant, les arbres phylogéniques (Figure 2) encodent également de l'information typologique, mais de manière implicite. En effet, bien que les reconstructions phylogéniques se basent le plus souvent sur des mesures lexicales, souvent le seul matériau disponible, l'on s'attend néanmoins à ce que des langues historiquement proches partagent aussi des traits grammaticaux. Ce que nous procurent les arbres phylogéniques est donc une notion de similarité entre les langues, sans nécessairement connaître la nature exacte des traits grammaticaux qu'elles partagent. Ceux-ci pourront néanmoins être inférés par les méthodes d'apprentissage à partir des données annotées. En outre, un avantage supplémentaire des arbres phylogéniques par rapport aux traits typologiques est qu'il existe plus de méthodes pour créer des arbres phylogéniques automatiquement à partir de listes de mots que pour trouver des traits typologiques à partir de texte.

Apprendre des modèles de manière évolutive

Notre approche rompt avec les travaux précédents, dans la mesure où nous proposons d'apprendre simultanément les modèles d'analyse associés aux différentes langues en exploitant directement l'arbre phylogénique sous-jacent, et sans autre information typologique.

Les modèles d'analyse syntaxique, appris automatiquement sur des corpus annotés, sont des représentations très abstraites et sommaires de la grammaire des langues, mais l'on peut supposer

1. Zeman D., Nivre J., Abrams M. et al. 2020, *Universal dependencies 2.7*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
2. Voir la version 2.7, parue en novembre 2020.
3. Naseem T., Barzilay R. et Globerson A. 2012, Selective sharing for multi-lingual dependency parsing, in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers* - Volume 1: 629–637, Stroudsburg, PA, USA.
4. Aufrant L., Wisniewski G. et Yvon F. 2016, Zero-resource Dependency Parsing: Boosting Delexicalized Cross-lingual Transfer with Linguistic Knowledge, in *Proceedings of the 26th International Conference on Computational Linguistics*, pp119–130, Osaka, Japan.

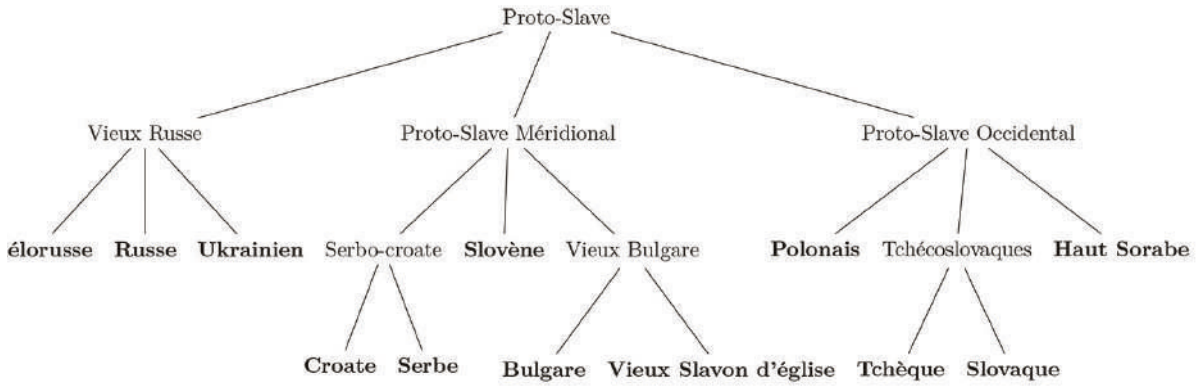

Figure 2 - Exemple d'arbre phylogénique. Rameau slave de l'arbre phylogénique utilisé pour nos travaux. Les noms des nœuds intermédiaires sont indicatifs, seules la structure de l'arbre et la place des feuilles (en gras) sont importantes. Le pluriel tchécoslovaque réfère aux langues du groupe tchécoslovaque et non pas à la norme écrite utilisée dans les années 1920 et 1930.

qu'ils évoluent d'une manière similaire à ces langues. Par exemple, si l'on avait des corpus pour les langues romanes représentant leur état tous les cinquante ans en commençant en l'an zéro et jusqu'aujourd'hui, l'on pourrait s'attendre à ce que des modèles d'analyse, appris sur chacun de ces corpus, forment une structure similaire à celle proposée par les linguistes historiques pour les langues romanes elles-mêmes. L'on s'attendrait entre autres à ce que des modèles représentant des corpus proches en terme de dates et de variétés (picard de 1150 et normand de 1300), soient plus proches que des modèles représentant des variétés très différentes (portugais de 1900 et roumain de 1300).

L'on pourrait alors faire évoluer les modèles lentement pour suivre l'évolution des langues au plus près, gardant les traits hérités d'époques antérieures et remplaçant les archaïsmes par leur formes modernes.

Malheureusement, nous n'avons pas accès à ce genre de données, entre autres parce que *Universal Dependencies* ne dispose très majoritairement d'annotations que pour des langues encore vivantes, beaucoup de langues n'ayant d'ailleurs commencé à être écrites que très récemment. Cependant, l'on peut utiliser un mélange de données venant de langues apparentées pour simuler leur langue mère. Plus précisément, l'on considérera comme exemple de la langue mère l'ensemble des annotations associées à ces descendants. Bien que cette approche soit très simpliste et qu'elle ait peu de chance de représenter fidèlement la grammaire de la langue mère, c'est peut-être là une de ses forces. En effet, l'objectif final étant d'apprendre des modèles d'analyse pour les langues modernes[5], il paraît pertinent de confronter les modèles à des données similaires à celles qu'ils auront à traiter à la fin et ce, dès le début de l'apprentissage.

Nous proposons alors de faire « évoluer » des modèles d'analyse en parallèle pour plusieurs langues historiquement apparentées (Figure 3). L'idée est qu'en entraînant un modèle d'analyse sur des données venant de plusieurs langues proches, on l'incite à se concentrer sur les caractéristiques communes à ces langues, dans notre cas, sur les traits syntaxiques partagés. L'on peut ensuite faire des copies du modèle appris pour imiter la langue mère et les passer aux rameaux descendants pour continuer l'apprentissage en partant d'un modèle déjà plus robuste. En s'enfonçant de plus en plus profondément dans l'arbre, les modèles se spécialisent de plus en plus. Appris pour une famille de langues au départ, ils se spécialisent ensuite sur un rameau et, enfin, sur une seule langue comme représenté dans la figure 2. Cependant, comme ils ont vu des données d'autres langues plus ou moins proches de leur langue cible, ils ont emmagasiné de l'information syntaxique qui, tout en étant utile à l'analyse de la langue cible, n'aurait pas nécessairement été disponible dans le seul corpus de la langue cible.

Résultats empiriques

Nous avons mené des expériences sur un ensemble de corpus annotés avec des arbres de dépendances rendus disponibles par le projet *Universal Dependencies*. Nous avons entraîné des modèles à reconstruire les arbres en dépendances à partir de mots de la phrase et de leurs attributs morphologiques (genre, personne, cas, temps, mode, etc.). Les modèles étaient soit entraînés ensemble le long d'un arbre phylogénique représentant les relations des différentes familles et sous familles de langues, soit entraînés indépendamment directement pour chacune des langues disponibles.

Les modèles (linéaires et neuraux) entraînés dans un arbre phylogénique représentant une partie de l'histoire évolutive des langues ont montré de meilleurs performances que des modèles entraînés avec les mêmes données mais de manière indépendante pour chaque langue. Les performances des modèles phylogéniques sont d'autant plus remarquables que les langues en question sont peu dotées et que leur famille linguistique est bien fournie en terme de langues.

Un exemple de ce phénomène est donné par le haut sorabe, une langue slave parlée dans l'est de l'Allemagne par environ 15 000 personnes. Bien qu'elle soit très peu dotée (une trentaine de phrases d'entraînement chez *Universal Dependencies*), sa proximité avec le polonais, le tchèque et le slovaque fait que l'on peut entraîner des modèles fiables en se basant sur les données disponibles pour ses langues sœurs.

En outre, les modèles entraînés pour un ensemble de langues proches peuvent aussi servir de point de départ pour analyser des langues pour lesquelles nous n'avons pas de données d'apprentissage du tout.

5. Par « langues modernes », nous entendons des langues pour lesquelles nous avons des données. En ce sens, le latin, le gotique ou encore le sanskrit sont des langues modernes, car il existe une littérature dans ses langues qu'il est intéressant d'analyser avec des outils numériques modernes.

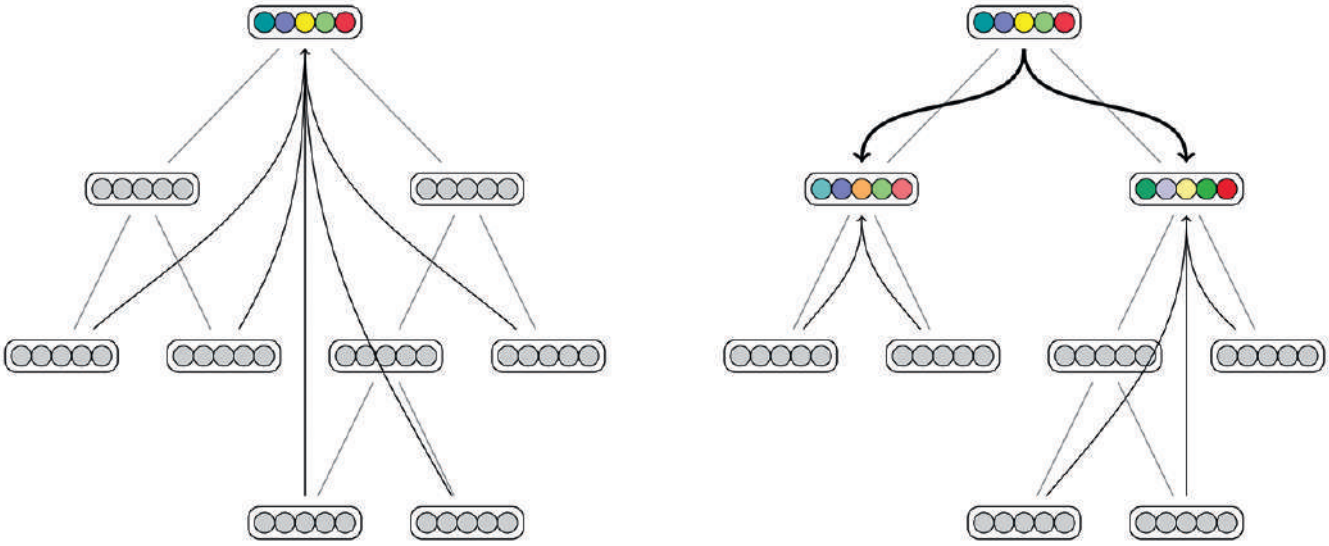

Figure 3. Représentation de la méthode d'apprentissage phylogénique. Les données de toutes les feuilles sont utilisées pour entraîner le modèle de la racine (gauche). Le modèle est ensuite copié et passé à ses descendants directs où l'apprentissage reprend (droite). Chaque nœud n'a accès qu'aux données de ses propres descendants.

Conclusion

Nos travaux montrent que les modèles utilisés pour le TAL, en l'occurrence des modèles d'analyse syntaxique, peuvent tirer profit de l'information historique sur l'évolution des langues, représentée ici par un arbre phylogénique, pour apprendre mieux des données disponibles en partageant les ressources existantes pour différentes langues.

Une des limitations de l'approche présentée ici est qu'elle utilise un arbre phylogénique comme source d'information historique. Or, les arbres phylogéniques ne sont pas adaptés à la représentation des transferts horizontaux[6] que l'on peut observer dans les aires linguistiques, par exemple. Les traits acquis par ces biais sont toutefois présents dans les langues et nous devons les y analyser. Il nous faudra donc envisager des structures plus expressives que les arbres, comme les graphes dirigés acycliques[7], si nous voulons avoir des représentions plus fidèles de l'évolution des langues. De plus, il nous faudra des méthodes pour créer automatiquement ces nouvelles structures à partir de données textuelles si nous voulons les utiliser à plus grande échelle et pour les langues sous dotées.

Nous invitons le lecteur intéressé à se référer à Dehouck et Denis[8] pour plus de détails concernant les algorithmes utilisés et pour des résultats complets.

contact&info
▶ Mathieu Dehouck
Lattice
mathieubmddehouck@mailoo.org
▶ Pascal Denis
CRIStAL
pascal.denis@inria.fr

6. Transferts horizontaux : emprunts de mots ou de structures grammaticales à des langues par contact, sans forcément que celles-ci ne partagent un ancêtre commun.
7. Graphes dirigés acycliques : structure dans laquelle les nœuds peuvent avoir plusieurs parents directs. Dans un arbre, les nœuds n'ont qu'un parent.
8. Dehouck M. et Denis P. 2019, Phylogenic multi-lingual dependency parsing, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pp 192–203, Minneapolis, Minnesota.