# L'arbre des langues indo-européennes

De nouveaux outils statistiques ont permis de construire un arbre généalogique de la famille des langues indo-européennes.


© DR

**Robin Ryder,** après avoir fait sa thèse à Oxford sur les modèles statistiques de diversification des langues, est aujourd'hui post-doctorant au laboratoire de statistiques du Centre de recherche en économie et statistique, à Paris.

**Pourquoi un mathématicien s'intéresse-t-il aux langues indo-européennes ?**

**R.R.** Cette famille de plusieurs centaines de langues, dont des langues qui ont disparu, comme le latin, a donné naissance aux langues romanes, dont le français. Les linguistes cherchent à reconstruire un arbre de parenté de toutes ces langues, ce qui nécessite d'analyser leurs points communs et leurs différences. C'est là que les statistiques sont utiles.

**Qu'apportent-elles à l'étude de l'arbre de ces langues ?**

**R.R.** Elles aident à dater les séparations entre les différentes langues et à estimer le degré d'incertitude des regroupements par branches. Elles permettent aussi de traiter de grands ensembles de données, difficiles à analyser à la main. Les linguistes avaient déjà une bonne idée de la structure de l'arbre, et quelques indications pour les dates. Par exemple, si deux langues cousines ont le même mot pour « charrue », cela peut signifier que la langue ancestrale commune était parlée par une population qui connaissait la charrue. En comparant les langues et munis de ce type d'indication, les linguistes ont établi des arbres, mais la chronologie restait incertaine. Dans les années 1950, les premières tentatives d'utiliser les statistiques ont été des échecs qui ont discrédité les méthodes mathématiques de datation.

**Comment dépasser les défauts de ces méthodes ?**

**R.R.** Celles-ci supposaient notamment que les langues se diversifient à un taux constant. En appliquant aux langues les méthodes de reconstruction utilisées en biologie moléculaire, qui se passent en partie de cette hypothèse, des premiers résultats encourageants ont été publiés en 2003. Avec l'aide de linguistes, nous avons décidé de construire un modèle spécifique de la diversification des langues qui traite de nombreuses langues simultanément. Nos données sont constituées d'une centaine de mots du vocabulaire de base, des mots comme « arbre » ou « animal », qui existent dans presque toutes les langues.

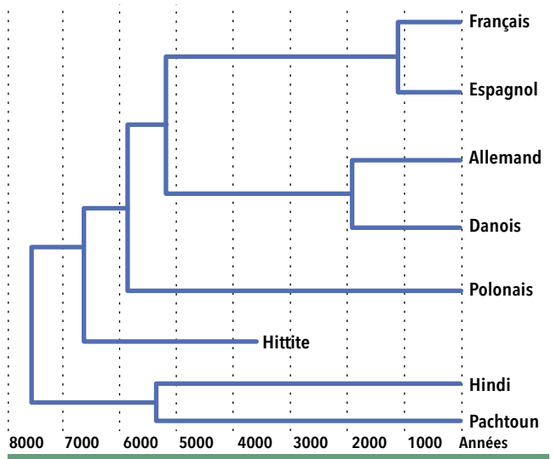
**Comment reconstruisez-vous l'arbre phylogénétique de ces langues ?**

**R.R.** Le premier travail consiste à transformer ces données linguistiques en données que nous pouvons traiter statistiquement. Pour cela, on cherche à déterminer si deux mots proches ont le même sens et s'ils ont une origine commune (des mots « cognats »). En établissant ces classes de cognats, les données lexicales sont codées par une matrice avec des 0 et des 1. Munis de ces données, nous tentons de calculer la fonction qui représente la densité de probabilité des différents paramètres du problème (taux de diversification, âge des nœuds de l'arbre, etc.). Autrement dit, nous cherchons quels sont les paramètres les plus vraisemblables. Pour explorer cette fonction, nous avons utilisé une méthode statistique récente, dites de chaînes de Markov Monte-Carlo (lire l'encadré ci-dessus).

**Pouvez-vous vérifier la fiabilité des datations ?**

**R.R.** Avant de mettre en œuvre le modèle sur les vraies données, nous l'avons testé avec des données simulées et sur des données pour lesquelles la réponse était connue par ailleurs. Les bons résultats nous ont permis de vérifier la cohérence du modèle et la fiabilité de nos estimations. De plus, nous obtenons des résultats comparables avec deux jeux de données indépendants, l'un de 24 langues anciennes et l'autre de 87 langues modernes. Ce que nous obtenons est un arbre avec des datations et les barres d'incertitude associées [1]. Nous avons ainsi établi que l'âge de la racine de l'ancêtre commun à toutes les langues était compris entre 7100 et 9800 ans avant aujourd'hui, avec une probabilité de 95 %. Cette période coïncide avec l'essor de l'agriculture, ce qui expliquerait l'extension et le développement de ces langues. ■ **Propos recueillis par Philippe Pajot**

[1] R. Ryder et G. Nicholls, *Journal of the Royal Statistical Society : Series C (Applied Statistics),* 2010.




Cette arbre représente une généalogie probable des langues reconstruite par la méthode statistique avec la datation des nœuds intermédiaires (les barres d'incertitude ne sont pas indiquées).

### L'EXPLORATION DES GRANDS ESPACES STATISTIQUES

L'espace statistique à explorer dans le cas de l'arbre des langues indo-européennes est immense. Pour 87 langues, il existe environ $2^{163}$ arbres différents. Et il faut multiplier ce nombre par tous les âges possibles pour tous les nœuds internes de l'arbre, soit 86 âges qui prennent des valeurs continues sur un intervalle. Le calcul de la fonction qui représente la densité de probabilité des différents paramètres du problème sur tout cet espace est impossible dans des temps raisonnables. La méthode de chaînes de Markov Monte-Carlo vise à obtenir un échantillonnage statistique de cette fonction. Il s'agit d'une marche aléatoire où l'on explore toutes les régions intéressantes de l'espace des paramètres, celles où les densités de probabilités sont les plus élevées. Au final, le calcul de la fonction converge en environ 12 heures.

### sur le web

**www.ethnologue.com**
Un site encyclopédique qui regroupe des milliers de langues, notamment par familles.

**http://en.wikipedia.org/wiki/Markov_chain_Monte_Carlo**
La page Wikipédia anglaise sur les méthodes de chaînes de Markov Monte-Carlo.

### Chronologie

**XVII^e^ et XVIII^e^ siècles :** les linguistes remarquent des similarités entre les langues et découvrent la famille des langues indo-européennes.
**1871 :** Charles Darwin note que l'évolution des espèces et celle des langues procèdent de mécanismes analogues.
**1950 :** Morris Swadesh fonde la glottochronologie, méthode pour étudier les relations chronologiques entre les langues.
**1962 :** Knut Bergsland et Hans Vogt montrent que les datations issues de la glottochronologie sont en contradiction avec les données historiques.
**2003 :** Russell Gray et Quentin Atkinson, de l'université d'Auckland, en Nouvelle-Zélande, reprennent le problème de la datation en se fondant sur des modèles de biologie moléculaire.
**2010 :** à l'université d'Oxford, Robin Ryder et Geoff Nicholls développent des modèles statistiques spécifiques aux langues et estiment l'âge de la plupart des embranchements de l'arbre des langues indo-européennes.