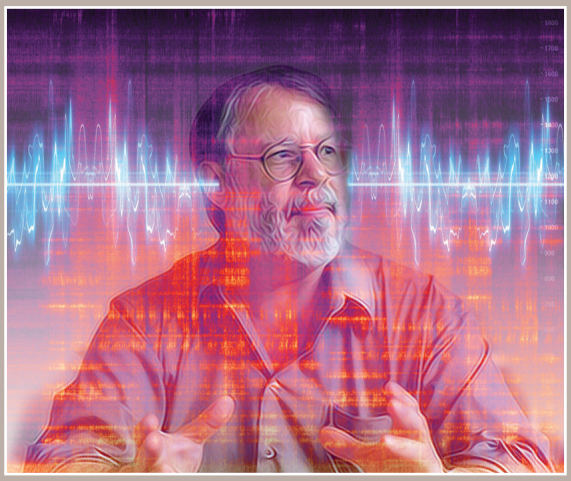
Mark Liberman

Les raisons du (quasi-) succès du traitement automatique de la parole



Prisme N°32
Décembre 2017

La Fondation et le Centre Cournot

La Fondation et le Centre Cournot

Association reconnue d'intérêt général, le Centre Cournot contribue à développer et à rendre plus accessibles les théories des sciences économiques, sociales, politiques et leur épistémologie. Comme le Centre, la Fondation poursuit un travail de catalyse dans la filiation d'Augustin Cournot, accélérant ou sélectionnant des formulations théoriques concurrentes. Sous égide de la Fondation de France, la Fondation met ainsi en perspective le basculement probabiliste qui atteint progressivement les disciplines et promeut les recherches qui élaborent et utilisent les probabilités.

Les raisons du (quasi-) succès du traitement automatique de la parole[1]

Mark Liberman

Prisme N°32

Décembre 2017

[1] Traduction de l'anglais par Nathalie Ferron.

Réimpression, juin 2018

Le Centre remercie Edouard Geoffrois pour sa relecture attentive.

En quoi le traitement automatique de la parole – reconnaissance et synthèse, traduction automatique, extraction d'information, systèmes de questions-réponses, etc. – est-il un succès ? Que peuvent apporter aux autres domaines les avancées réalisées dans les disciplines qu'il recouvre ? Je voudrais commencer par rappeler comment fonctionne le traitement de la parole à l'aide de quelques exemples.

Prenons l'exemple des *smartphone*s. La plupart des téléphones sont désormais équipés d'une fonction vocale de questions-réponses : Siri sur les iPhones, OK Google sur Android et une fonction similaire pour les téléphones Windows. Ce matin, j'ai donc allumé mon téléphone et j'ai dit :

« OK Google, comment dit-on 'chien' en français ? »

Je m'attendais bien à ce que la transcription soit correcte, mais j'eus la surprise de constater que non seulement la transcription était correcte mais qu'en plus, grâce à la synthèse vocale, on me donnait la version orale du mot.

Assez admiratif, je posai une nouvelle question :

« OK Google, combien font 15 degrés Celsius en Fahrenheit ? »

J'obtins une transcription correcte plus une réponse immédiate :

« 15 degrés Celsius font 59 degrés Fahrenheit ».

La réponse m'est parvenue sous forme de texte écrit, non sous forme de paroles, mais mon assistant personnel aurait parfaitement pu en faire la synthèse vocale. Je posai une nouvelle question :

« OK Google, quel est le nom du journal étudiant de l'université de Pennsylvanie ? »

En même temps qu'une transcription correcte, j'obtins une page comportant une liste de liens vers *The Daily Pennsylvanian*, qui est le nom du journal en question.

« OK Google, pense-bête : acheter de l'essuie-tout. »

Transcription :

« Pense-bête : acheter de l'essuie-tout. »

Transcription assortie d'un courriel dont j'étais l'expéditeur, m'informant que je devais acheter de l'essuie-tout.

Je décidai qu'il était temps de le mettre en échec, de lui demander quelque chose qu'il ne saurait pas faire. J'avais sous la main un manuel de bibliothèque graphique, *ggplot2*, une extension de R, écrit par un certain Hadley Wickham. Je posai donc ma nouvelle question :

> « OK Google, quelle est la date de parution du livre de Hadley Wickham, *ggplot2* ? »

Je pensais bien que jamais je n'aurais la réponse, et de fait, la transcription suivante s'afficha sur mon écran :

> « Quelle est la date de parution du livre de Hadley Wickham, *ggplot2* ? »

Je me suis quand même demandé comment ce titre avait bien pu intégrer son lexique. Il m'a ensuite proposé une page comportant les résultats de ma recherche : le site d'Amazon arrivait en tête avec le titre du livre. Comme je n'avais pas réussi à le prendre en défaut, j'essayai autre chose :

> « OK Google, comment dit-on 'chien' en langue haoussa ? »

Réponse :

> « Voici votre traduction » (sous forme vocale)

Après quoi j'ai été redirigé sur *Google translate* avec mon mot en haoussa. C'était presque inquiétant ! Je renonçai à faire dérailler OK Google même si je suis certain que j'aurais pu y parvenir. Dans un environnement bruyant par exemple, il commencerait sans aucun doute à donner des signes de faiblesse.

Je suis ensuite allé sur *Google translate*, et j'ai fait un copier-coller d'un passage tiré du site du Centre Cournot dans sa version française :

> « Le Centre Cournot est une association soutenue par la Fondation Cournot, placée sous l'égide de la Fondation de France. Elle porte le nom du mathématicien et philosophe

> **franc-comtois** Augustin Cournot (1801–1877), reconnu de longue date comme un pionnier de la **discipline économique**. »

Traduction de Google :

> « The Cournot Centre is an association supported by the Cournot Foundation, under the aegis of the Fondation de France. It is named after the mathematician and philosopher **Franc-Comtois** Augustin Cournot (1801-1877), long recognized as a pioneer of **economic discipline**. »

Cette fois, il y avait quelques erreurs. Par exemple, *Google translate* ne sait pas que « franc-comtois » est un adjectif. Il dit aussi que Cournot est reconnu comme un pionnier en matière de « discipline économique », ce qui ne correspond pas à ce que l'on entend en français par « discipline économique ». La traduction exacte serait « *economics* » ; « *economic discipline* » renvoie en effet à la notion d'austérité en matière économique.

Site du Centre :

> « Le Centre n'est pas un laboratoire de recherche, il n'est pas non plus un centre de réflexion. **Il** jouit de l'indépendance singulière d'un catalyseur. »

Traduction Google :

> « The Centre is not a research laboratory, it is not a think tank. **He** enjoys the singular independence of a catalyst. »

On remarquera dans ce segment que le pronom « il », qui en anglais devrait se traduire par « it » dans la mesure où il a pour antécédent « le Centre », est traduit par « ***he*** *enjoys the singular independence of a catalyst...* ».

De même :

> « Pour qu'un débat ait lieu, il faut plus que de la connaissance et de la compréhension. Il faut des préférences, des croyances, des désirs, des objectifs… **C'est en pratique**

> **de cela seulement dont les débatteurs disposent et ils inventent ou ils adoptent les résultats qui leur conviennent**. »

Traduction Google :

> « To have a debate, it takes more than knowledge and understanding. It takes preferences, beliefs, desires, goals... **In practice this only with the debaters have and they invent or they adopt the results that suit them.** »

Dans ce dernier passage, *Google translate* s'en sort bien sur les deux premières phrases mais se trompe sur la dernière. Il lui arrive de se fourvoyer au moins de temps de temps.

Enfin, je suis en train de lire un roman policier dont le titre est *Le Dingue au Bistouri*[2]. Il commence ainsi :

> Il y a quatre choses que je déteste.
> Un : qu'on boive dans mon verre.
> Deux : qu'on se mouche dans un restaurant.
> Trois : qu'on me pose un lapin.
> […]

Traduction Google :

> There are four things I hate.
> A: we drink in my glass.
> Two: we will fly in a restaurant.
> Three: I get asked a rabbit.
> […]

La traduction que propose *Google translate* est erronée sur l'intégralité du passage. « Qu'on boive dans mon verre » se traduit par « *someone drinks from my*

[2] Khadra, Yasmina (1983), *Le dingue au bistouri,* Paris: Éditions Poche.

glass », or Google propose « *we drink in my glass* ». « Qu'on se mouche dans un restaurant » signifie qu'on se débarrasse les narines de leurs mucosités, mais le traducteur de Google confond le verbe « se moucher » avec le nom de l'insecte, ce qui donne une phrase dépourvue de sens, « *we will fly in a restaurant* ». Quant à l'expression idiomatique, « poser un lapin », Google n'en connaît apparemment pas le sens puisqu'il la traduit par « I get asked a rabbit ».

Le traducteur de Google a donc ses failles, mais il faut quand même aller chercher assez loin pour le prendre en défaut, du moins pour des langues répandues comme le français ou l'anglais. Par souci d'équité, j'ai voulu donner sa chance au traducteur de Bing. Il s'en est tiré plutôt moins bien, pas mieux en tout cas.

Revenons à notre sujet principal, au fait que le traitement automatique de la langue naturelle est un succès, quasiment. Quelles en sont les raisons ? Quelle en est l'histoire ? Certaines raisons sont évidentes : il existe une sorte d'univers numérique parallèle qui reflète de mieux en mieux la réalité sous forme de flux et de bits. Toute société repose avant tout sur la communication, or qu'est-ce que la communication sinon essentiellement du texte ou de la parole (qui n'est autre que du texte sous une forme un peu particulière), de plus en plus souvent sous forme numérique ? Certaines propriétés élémentaires d'un texte (comme les mots qui le composent par exemple) constituent un bon indicateur de son contenu. Nous disposons aujourd'hui d'un arsenal numérique – réseaux, ordinateurs, téléphones… – toujours plus vaste, plus rapide et moins cher, et de langages de programmation plus performants qui facilitent l'extraction de contenus au sein du flux textuel propre à cet univers numérique parallèle.

Qui, du contenu ou de la communication, doit avoir la primauté ? Le débat n'est pas nouveau. Du moins « le contenu de la communication » est-il l'élément moteur qui réconcilie les deux parties. Une nouvelle niche évolutionnaire donne à ces nouvelles formes de vie – OK Google, Siri, et autres –la capacité, la possibilité et de bonnes raisons d'occuper leur niche écologique tout en faisant bénéficier l'environnement du produit de leur propre digestion. C'est précisément l'une des raisons pour lesquelles le traitement automatique des langues naturelles est de plus en plus performant : il crée des possibilités.

Une autre raison du quasi-succès du traitement automatique de la langue tient naturellement aux progrès de l'apprentissage automatique, c'est-à-dire essentiellement aux statistiques appliquées et aux capacités des ordinateurs à les mettre en œuvre. Toutes sortes d'acronymes et de néologismes ont été créées pour rendre compte de nouvelles techniques, comme celui des unités LSTM (Long Short-Term Memory), la mémoire de court-terme persistante, qui constitue des éléments simples des réseaux de neurone récurrents ou comme celui de « réseau neuronal profond ». Ces notions sont souvent simples d'un point de vue conceptuel mais complexes du point de vue mathématique, et très utiles à la résolution de divers problèmes relatifs à l'analyse de la parole et de la langue.

Des progrès considérables ont été accomplis dans ce domaine au cours des dernières décennies.

Il est une troisième raison à mes yeux plus importante encore que les deux précédentes, qui tient à un changement culturel survenu il y a un demi-siècle. C'est de ce changement dont je voudrais parler à présent.

Mon propos s'appuie sur une présentation que j'ai faite en 2015 au cours d'un atelier qui s'est tenu à la ***National Academy of Sciences*** des Etats-Unis sur le thème des problèmes statistiques liés à l'évaluation et à la reproductibilité des résultats de recherche[3]. Cet atelier a mis au jour de réelles inquiétudes et c'était son but dans la mesure où la crédibilité de nombreux domaines de recherche scientifique se trouve aujourd'hui remise en question. Les exemples sont légions, je n'en citerai que deux : un article célèbre de John Ioannidis, intitulé « ***Why Most Published Research Findings are False*** »[4], qui remonte à 2005 mais n'a rien perdu de sa validité et plus récemment, un article intitulé « ***Amid a Sea of False Findings, the NIH Tries Reform*** »[5]

[3] Michelle Schwalbe, Rapporteur ; Committee on Applied and Theoretical Statistics ; Board on Mathematical Sciences and Their Applications ; Division on Engineering and Physical Sciences ; National Academies of Sciences, Engineering, and Medicine (2016), *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop*, Washington, D.C. ; The National Academies Press.

[4] Ioannidis, John (2005), « Why most published research findings are false », *PLoS Medicine*, 2(8), August, e124.

[5] Voosen, Paul (2015), « Amid a Sea of False Findings, the NIH Tries Reform », *The Chronicle of Higher Education*, 16 March.

paru dans la revue *Chronicle of Higher Education*. L'auteur y cite le Dr Francis Collins, directeur du *National Institutes of Health* (le plus grand centre de recherche biomédicale biologique au monde, basé aux États-Unis) rapportant que des chercheurs spécialistes de la sclérose latérale amyotrophique (SLA) ont dans le cadre de leur recherche d'un traitement, tenté de reproduire des études portant sur plus de 70 médicaments prometteurs. « Aucune de ces études n'étaient reproductibles, » précise le Dr Collins. « Aucune, et pourtant on était déjà passé aux essais cliniques sur des personnes pour certains d'entre eux… ». Beaucoup de personnes s'intéressent à ce problème à travers le monde. Des psycholinguistes de l'École normale supérieure (ENS) ont organisé une série d'ateliers sur la reproductibilité en psychologie, mais le problème ne concerne pas uniquement la psychologie, il touche aussi la recherche biomédicale et bien d'autres domaines. C'est pourquoi je vais maintenant vous raconter l'histoire d'une crise de crédibilité qui a touché un autre domaine de recherche il y a 50 ans.

Il était une fois un chercheur des laboratoires Bell qui s'appelait John Pierce. Il dirigeait l'équipe qui a conçu le premier transistor et suivi le développement du premier satellite de communication. John Pierce n'avait aucun problème de crédibilité. On le voit dans la photographie reproduite à gauche. Il a la posture typique des ingénieurs des années 1950 : vêtu d'un costume-cravate, il se tient devant un appareil analogique constellé de cadrans et de boutons et très compliqué. En 1966, Pierce était à la tête d'un comité – l'*Automatic Language Processing Advisory Committee*, plus connu sous l'acronyme ALPAC – chargé de rédiger un rapport sur la traduction automatique pour la

Image 1 : Portrait de John Pierce dans les années 1950

National Academy of Sciences. Le rapport de l'Alpac[6] soulignait que la traduction automatique ne donnait pas de résultats très satisfaisants pour l'heure et proposait, avec les précautions d'usage, la conclusion suivante : « Le comité n'est pas en mesure de juger quel doit être le montant annuel des dépenses consacrées à la recherche et au développement pour améliorer la qualité de la traduction automatique. Cependant, il recommande que ce montant soit employé dans un souci de réalisme pour soutenir des projets importants, réalistes et de relatif court terme. » (ALPAC, 1966, p. 33). Ce qui signifie en langage ordinaire : « Arrêtez de leur donner de l'argent ! »

Aux yeux du comité, la science devait dans des cas comme celui-ci précéder l'ingénierie, aussi évoqua-t-il plus ou moins explicitement les merveilleuses possibilités qu'offrait l'informatique à la recherche en linguistique. Cependant, les pourvoyeurs de fonds saisirent parfaitement le message et suite à la publication du rapport, le financement de la traduction automatique aux États-Unis fut réduit à néant pour les vingt années qui suivirent. Pierce avait le même point de vue sur la reconnaissance automatique de la parole que sur la traduction automatique. En 1969, il écrivit au *Journal of the Acoustical Society of America* une lettre intitulée « Où va la reconnaissance automatique de la parole ? »[7], dans laquelle il donnait à entendre son opinion en des termes moins diplomatiques :

> *...il est proprement impossible d'envisager une machine à écrire phonétique tant qu'on n'en aura pas inventé une qui soit dotée d'une intelligence et d'une connaissance de la langue comparables à celles d'un locuteur natif. […] Au lieu de se comporter en scientifiques, la plupart des spécialistes de la reconnaissance* [c'est-à-dire des chercheurs en reconnaissance automatique de la parole] *agissent en savants fous ou en*

[6] ALPAC (1966), *Language and Machines: Computers in Translation and Linguistics.* A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council, Publication 1416.

[7] Pierce, John (1969), « Whither Speech Recognition? », *The Journal of the Acoustical Society of America,* 46(4), 1049–1051(L).

ingénieurs irresponsables. En règle générale, ils se croient en mesure de résoudre « le problème », soit en vertu d'une inspiration qui leur est propre (comme dans le cas du « savant fou », source de son propre savoir), soit parce qu'ils prennent pour argent comptant des règles, des modèles et des informations sans les avoir testés (l'approche de l'ingénieur irresponsable). […]

Ils [...] élaborent ou programment des systèmes très sophistiqués qui au mieux ne font pas grand-chose, au pire et pour une raison obscure, ne fonctionnent pas du tout. On investit beaucoup de temps et d'argent sans aboutir à aucune connaissance simple, claire et sûre. Leur travail leur fait vivre de belles expériences, certes, mais il ne s'agit pas d'expériences scientifiques.

Plus loin, il dit encore :

Nous pouvons sans risque affirmer que la reconnaissance de la parole attire les financements, de la même manière peut-être que d'autres projets comme la transformation de l'eau en carburant, l'extraction de l'or dans la mer, la guérison du cancer, ou l'exploration de la lune. Ce n'est pas avec des projets visant à réduire le prix du savon de 10% qu'on attire facilement l'argent. [Ce qui est sans doute faux empiriquement du point de vue économique : en proposant une réduction de 10% du prix du savon, on trouverait sans doute des investisseurs…] *Pour duper les acheteurs, il faut user de tromperie et de séduction.*

Dans le domaine de la reconnaissance de la parole, il semble évident que les bénéficiaires des fonds sont tout autant aveuglés par leurs mirages qu'ils n'aveuglent ceux qui les leur donnent. On peut éventuellement éprouver quelque pitié

> *pour ces travailleurs qu'il est impossible au demeurant de respecter.* (Pierce, 1966, pp. 1049-1051)

Pierce ne mâche pas ses mots. Bientôt, plusieurs sommités du domaine tout juste naissant d'une certaine forme d'intelligence artificielle objectèrent que le problème tenait au fait que les « savants fous » et les « ingénieurs irresponsables » en question ne connaissaient rien au langage Lisp (traitement de listes), à l'intelligence artificielle ou encore au calcul des prédicats du premier ordre appliqué aux problèmes de reconnaissance de formes. Ils persuadèrent donc la DARPA (Defense Advanced Research Projects Agency[8]) d'investir dans un programme visant à appliquer l'intelligence artificielle au problème de la reconnaissance vocale. Ce programme fondé sur l'utilisation d'une forme classique d'IA – la logique appliquée – avait pour but de permettre la compréhension de messages parlés avec la même facilité quasiment qu'un locuteur natif. Cette méthode s'appuyait largement sur des idées a *priori* concernant ce qui était dit.

L'un des programmes permettait de jouer aux échecs contre un robot en annonçant ses coups à voix haute plutôt qu'en utilisant un échiquier ou une interface graphique. On pouvait par exemple dire « tour deux cases vers la gauche » et la tour avançait si vous aviez été compris. Au cours d'une démonstration du projet, un des représentants de la DARPA, que la lecture de la prose de Pierce avait rendu sceptique, entreprit de jouer une partie d'échecs avec la machine en s'adressant à elle par des aboiements. Tandis qu'il faisait ses « ouah, ouah », la machine, qui avait une idée de l'espace sémantique possible, faisait exécuter des mouvements plausibles aux différentes pièces. Elle optait pour le coup qui correspondait le mieux à ce qu'on lui avait donné à entendre. Le projet fut bien entendu jugé non concluant et les financements stoppés prématurément au bout de trois années seulement. Il fut même envisagé de tout arrêter. Entre 1975 et 1986, les recherches concernant la traduction automatique et la reconnaissance automatique de la parole ne reçurent quasiment aucun financement aux États-Unis.

[8] La DARPA est une agence du ministère de la défense des Etats-Unis en charge du développement de techniques innovantes à usage militaire. En 2017, son budget était de 2,7 milliards d'euros. Elle correspond, toute proportion gardée, à la Délégation générale à l'armement en France.

John Pierce n'était pas seul à penser qu'il ne valait pas la peine d'investir dans ce domaine. De nombreux directeurs de recherche bien informés, et peut-être même la plupart d'entre eux, étaient tout aussi sceptiques sur les perspectives de ce champ de la science appliquée. Lorsque je suis entré dans les laboratoires Bell en 1975, en matière de reconnaissance de nombres isolés ou éventuellement connectés, seuls étaient admis les projets les plus limités, les plus timides et les plus modestes, même s'ils étaient considérés comme probablement voués à l'échec. Pourtant, au même moment, beaucoup de gens étaient convaincus qu'on avait besoin du traitement automatique de la langue naturelle ou du moins qu'on en aurait besoin à l'avenir, ils pensaient aussi que sur le principe la chose était faisable. 1985 marqua un tournant critique et le débat fut engagé au sein de certains cercles sur l'opportunité pour la DARPA de relancer la recherche sur ces technologies.

J'aurais peut-être dû rappeler que la DARPA n'est pas une agence ordinaire, et cela est vrai non seulement à l'échelle des États-Unis mais sans doute aussi à l'échelle mondiale ; en effet, son budget est très important or quasiment aucune des personnes qu'elle emploie ne l'est de façon permanente. Tous les chefs de projets et leurs équipes sont temporairement détachés de leurs postes à l'université ou dans les instances gouvernementales; quant au personnel administratif, il est fourni par des sous-traitants locaux pour une durée limitée et pour un projet conçu pour durer un certain temps avant de disparaître.

Charles Wayne, qui était en poste à la DARPA et venait de l'agence nationale pour la sécurité (*National Security Agency*) eut une idée. Il rejoignit la DARPA pour diriger un programme sur les technologies de la langue en 1985 et surmonter le scepticisme et les réactions négatives de ses supérieurs. Il prit garde qu'on ne puisse lui reprocher d'user de « tromperie et de séduction » en intégrant à son projet un système d'évaluation quantitative objectif et clairement défini prévu pour être appliqué par un tiers neutre – le National Institute of Standards and Technology (NIST) – à des bases de données partagées et non publiées. Le fait que les participants au projet eussent l'obligation de présenter leurs méthodes au bailleur de fonds et aux autres participants au moment de la publication des résultats de l'évaluation garantissait que les connaissances acquises seraient simples, claires et sûres, pour reprendre les termes utilisés par Pierce.

Pour parvenir à ses fins, il avait besoin de données publiées, dont une partie serait bien sûr mise de côté pour permettre de faire des tests, et d'un système d'évaluation clairement défini. Il s'assura donc le concours de David Pallett qui travaillait au *National Bureau of* Standards, aujourd'hui connu sous le nom de *National Institute of Standards and Technology*. Pallet se pencha sur la question et rédigea en 1985 un article pour le *Journal of Research of the National Bureau of Standards*[9] sur l'évaluation des performances des systèmes de reconnaissance automatique de la parole dans lequel on pouvait lire la chose suivante :

> *Il n'existe pas pour le moment de tests concluants permettant de décrire de manière exhaustive un système de reconnaissance automatique de la parole ni d'évaluer ses performances* [par quoi il entendait qu'il existe une grande diversité de conditions d'écoute, de discours, de locuteurs, de thèmes, etc.]. *Cependant, il est possible de concevoir et d'effectuer des tests d'évaluation des performances en puisant dans des bases de données orales largement accessibles, en utilisant des procédures de test similaires à celles que l'on utilise dans d'autres domaines et qui sont décrites avec précision. Ces tests permettent d'obtenir des données de références précieuses et ont un pouvoir prédictif intéressant malgré ses limites.* ***En revanche, les essais qui s'appuient sur des bases de données orales non ouvertes et dont les procédures et les résultats ne font pas l'objet de descriptions précises n'apportent que peu d'informations objectives sur les performances d'un système***. [en gras dans le texte] (Pallett, 1985, p. 371)

Peu avant la publication de cet article, George Doddington, ingénieur électricien travaillant chez Texas Instruments à l'époque avait attiré l'attention de certaines personnes à la DARPA, dont Charles Wayne, et de David Pallett au NIST.

[9] Pallett, David (1985), « Performance Assessment of Automatic Speech Recognizers », *Journal of Research of the National Bureau of Standards,* 90(5), septembre-octobre, pp. 371-387.

Une fois embauché chez Texas Instruments, la première chose qu'il entreprit de fabriquer fut une puce de synthèse de la parole par LPC[10], petite et bon marché, que l'on utilisait pour un jeu appelé ***Speak & Spell***, ***La Dictée magique***. Ce jeu électronique, qui coûtait environ 50 dollars, fut sans doute le premier jeu électronique proprement dit. On appuyait sur le bouton et on entendait : « épelle chat », on appuyait alors sur les touches correspondantes et la machine vous disait si la réponse était correcte ou non. Comme le composant mémoire coûtait très cher à l'époque – il était impossible de stocker des formes d'onde audio correspondant à des centaines, voire des milliers de mots – on utilisait la synthèse de la parole par LPC avec compression sur bande passante extrêmement faible. ***La Dictée magique*** eut un immense succès et rapporta beaucoup d'argent à ***Texas Instruments***. Interrogé sur ses projets à venir, G. Doddington annonça qu'il voulait travailler sur la reconnaissance automatique de la parole. Il commença par acheter un exemplaire de tous les programmes et de tous les appareils alors en vente sur le marché. La plupart provenaient du Japon car les entreprises de ce pays avaient continué de s'intéresser au problème. Ils garantissaient bien entendu une précision à 97,6%. Doddington créa sa propre base de données afin de les mettre à l'épreuve. Elle était constituée de nombres connectés prononcés dans leur langue par un large échantillon de locuteurs anglophones américains. Il testa chacun des programmes et des appareils qu'il s'était procuré à l'aide de sa banque de nombres connectés et évalua leur performance à partir d'une méthode de calcul simple qu'il avait élaborée. Il publia ses résultats[11] dans un article qu'il proposa à l'***IEEE Spectrum*** – un magazine haut de gamme largement diffusé et distribué à chacun des 30 à 40000 membres de l'IEEE[12].

Dans la mesure où il travaillait pour une entreprise, il devait soumettre son article à la direction pour avoir le droit de le publier, conformément à l'usage. L'article fut immédiatement rejeté au motif que l'information qu'il contenait était la propriété de l'entreprise. L'entreprise avait dépensé beaucoup d'argent pour évaluer le niveau de performance des meilleurs appareils de l'époque, aussi n'était-elle pas

[10] *Linear predictive coding,* codage par prédiction linéaire.

[11] Doddington, George, and T.B. Schalk (1981), « Speech recognition: Turning theory to practice », *IEEE Spectrum,* 18(9), septembre, pp. 26-32.

[12] Institute of Electrical and Electronics Engineers

prête à partager ces informations avec le reste du secteur. Doddington se passa d'autorisation et publia son article. Il fit même la couverture du magazine. Bien entendu, Richard Wiggins alors vice-président recherche chez Texas Instruments, reçut son magazine et découvrit la couverture. Il convoqua Doddington sur le champ et, rouge de colère, lui demanda en brandissant l'objet du délit ce que signifiait cette histoire. George s'assit, et lui répondit le sourire aux lèvres : « Je suppose qu'on peut appeler cela de l'insubordination caractérisée. La question est de savoir ce que vous allez en faire. » Bien entendu il obtint une promotion et se vit confier des projets de plus grande envergure. Cet exemple montre que cette démarche, visant à obtenir une évaluation comparative plausible pour une tâche limitée et pour un grand nombre de systèmes de reconnaissance automatique, frappa l'imagination de certains de ceux qui contribuaient à financer ce type de recherches.

La démarche de Doddington servit de modèle à ce qu'on appela par la suite « tâche partagée » et qui consiste en premier lieu à définir la tâche avec précision puis à élaborer un projet d'évaluation fondé sur la consultation répétée des chercheurs, un processus qui peut s'étaler dans le temps. Les financeurs du programme de recherche annonçaient ce qu'ils attendaient, les chercheurs venaient le trouver pour lui dire que la chose n'était pas possible ; ils proposaient de tenter autre chose et les échanges duraient ainsi jusqu'à ce qu'on parvienne à un accord. Un appel à candidatures était alors publié sur cette base pour lancer le projet. Ensuite, le NIST était sollicité pour développer un logiciel automatique d'évaluation, qui était également rendu public au début du projet. Ils commandaient la création de données d'apprentissage et de développement, également fournies en début de projet. Enfin, ils gardaient par-devers elle les données d'évaluation destinées aux évaluations publiques régulières.

La démarche ne plaisait pas à tout le monde. Beaucoup de personnes, parmi lesquelles John Pierce, demeuraient sceptiques à l'égard du projet : « Vous pouvez mesurer tout ce que vous voulez, ce n'est pas cela qui vous permettra de transformer l'eau en pétrole ». Les chercheurs, pour leur part, se sentaient remis en cause. Pour Richard Schwartz, qui travaillait alors chez Bolt Beranek and Newman (BBN) – qui avait été l'une des figures de proue des projets de recherche de la DARPA sur la compréhension du langage parlé entre 1972 et 1975, et qui fut pendant de longues années l'un des principaux responsables des travaux très importants et

internationalement reconnus menés par BBN sur la reconnaissance de la parole – c'était « comme si on était revenu au cours préparatoire : on vous dit exactement quoi faire et puis on vous évalue en permanence. »

Pourtant, cela a marché, pour la bonne raison que l'argent s'est mis à couler à flot, ce qui leur a permis de payer des gens pour travailler sur le projet. Certains problèmes trouvent leur solution par hasard alors qu'on cherche autre chose ; il y avait peu de chances que cela se produise avec la reconnaissance de la parole. Heureusement, dans la mesure où ils pouvaient mesurer les progrès accomplis au fil du temps, les bailleurs de fonds continuèrent d'accorder leurs financements, et ce fut une chance car le projet démarra en 1985–1986, et il fallut attendre 25 ans, voire plus, jusqu'à une date très récente en fait, pour parvenir à un résultat qui présentât un intérêt commercial, sous forme de produits contenant ces technologies et susceptibles d'être achetés par l'armée ou par le grand public.

Une autre raison, moins évidente celle-là, qui explique le succès du projet tient au fait qu'il accordait une place à l'utilisation d'algorithmes de recherche locale de type *hill climbing* grâce à sa méthode d'évaluation automatique et au caractère public du code d'évaluation. Une nouvelle méthode de travail s'imposa donc, qui fut pour bien des chercheurs une véritable révélation. Les mêmes chercheurs qui naguère rechignaient à être évalués deux fois par an se mirent à se tester eux-mêmes aussi souvent qu'ils parvenaient à réécrire leur code, toutes les heures, tous les jours, ou toutes les semaines.

Une dernière raison, encore moins visible, qui a pu permettre au projet de marcher, c'est qu'il fit éclore une culture fondée sur le partage des méthodes et des résultats à partir de données partagées et selon une méthode d'évaluation commune. Par son efficacité, cette culture exerça un tel attrait sur les chercheurs que nombre d'entre eux se joignirent au projet même sans financement. Ainsi, lors d'une des premières conférences sur la recherche d'information (Text Retrieval Conference ou TREC), financées par le Ministère américain de la défense, 40 laboratoires s'inscrivirent pour participer à l'évaluation et présenter leurs propres recherches alors que le ministère n'avait proposé de financer que quatre « acteurs » ou sites ayant un contrat de recherche dans le domaine. Le ministère prit alors conscience que le

partage des données et le fait de rendre publics les spécifications et le cadre d'évaluation permettaient de faire avancer la recherche sans payer de chercheurs.

La recherche sur la reconnaissance de formes en général et sur le traitement automatique de la parole et de la langue naturelle en particulier changea de nature. Lorsque tous les programmes doivent interpréter un même contenu ambigu, la résolution de ces ambigüités devient un jeu dans lequel le recours aux méthodes statistiques et probabilistes se trouve récompensé, ce qui a conduit au succès de l'apprentissage automatique. L'intelligence artificielle qui, dans les années 1970 et 1980, était de la logique appliquée, était devenue pour l'essentiel de la statistique appliquée. Avec la nouvelle génération, une petite part de logique se trouve peu à peu réintroduite, mais les changements introduits par l'IA furent vraiment majeurs.

Étant donné la nature de la parole et du langage, il est nécessaire d'adosser les méthodes statistiques à un ensemble de données d'apprentissage le plus large possible, ce qui renforce la valeur des données partagées dans la mesure où les quantités de données qu'un groupe peut se procurer collectivement sont généralement plus grandes que peut obtenir chaque individu. Dans ce jeu, le processus itératif d'apprentissage/évaluation a des vertus proprement addictives (je ne serais pas étonné en effet qu'il y ait ici libération de dopamine), mais il ne s'arrête pas là : en plus de rendre les chercheurs accros à l'évaluation, il leur permet de produire un savoir simple, clair et sûr qui incite le plus grand nombre à s'engager dans cette culture de la tâche partagée. Ces chercheurs ressemblent sans doute un peu aux joueurs de casino qui mettent leur argent dans des machines à sous et actionnent les manettes, sauf que le résultat en ce qui les concerne est plus productif.

Cette méthode de la tâche partagée est ainsi devenue le paradigme de la recherche en sciences numériques expérimentales, et pas seulement dans le domaine du traitement automatique de la parole et du langage. Elle est fondée sur la mise à disposition du public des données d'apprentissage et d'évaluation, sur une définition rigoureuse des critères d'évaluation et sur diverses techniques visant à éviter le sur-apprentissage, ce qui implique d'éviter très soigneusement de tester sur les données d'apprentissage, mais aussi de garder des données d'évaluation pour une évaluation vraiment indépendante. Ces méthodes sont à la fois statistiques et organisationnelles. La méthode de la tâche partagée peut s'appliquer à toutes sortes de domaines,

notamment à tout ce qui relève de l'analyse algorithmique du monde naturel. D'autres domaines peuvent être concernés, mais la méthode est particulièrement adaptée à l'interprétation des faits et des observations directes du monde.

Depuis 1985, des variantes de cette méthode ont été appliquées à des dizaines d'autres domaines : traduction automatique, identification du locuteur et de la langue parlée, analyse grammaticale, désambiguïsation sémantique, récupération et extraction d'information, synthèse, systèmes de questions-réponses, reconnaissance optique des caractères, analyse des sentiments, analyse d'images, analyse de vidéos, etc. jusqu'aux systèmes de navigation pour véhicules autonomes et à certains aspects de la robotique. L'expérience montre qu'en général les taux d'erreurs diminuent chaque année d'un pourcentage à peu près constant, ou autrement dit que la performance s'améliore exponentiellement, jusqu'à atteindre une asymptote qui dépend de la tâche et de la qualité des données d'apprentissage et d'évaluation.

On progresse en général par petites améliorations successives. Cela peut parfois être un peu démotivant. On assiste à une conférence ou à un atelier, on écoute des dizaines de présentations, les unes déploient des analyses conceptuelles de toute beauté, les autres déroulent le résultats de calculs extrêmement complexes, d'une programmation délicate, de semaines passées à faire mouliner des ordinateurs super rapides… tout cela pour un gain de performance dérisoire. Et pourtant il y a bien de quoi se réjouir : en effet, en s'ajoutant les uns aux autres, des progrès même minimes finissent par représenter quelque chose. Un progrès d'un pour cent est déjà significatif. Lorsque, en s'appuyant sur ce que l'on appelle les réseaux de neurones profonds, des chercheurs ont réussi à améliorer les performances de l'un des critères de reconnaissance de la parole en sorte que le taux d'erreur a diminué d'un tiers – pour passer de 20 à 14 pour cent par exemple – ce fut une victoire, le progrès le plus important depuis des décennies. Le partage des données joue un rôle essentiel et il se trouve souvent réutilisé dans des configurations inattendues. Séduction et tromperie ont de façon générale été éliminées. Bien sûr, la tentation a tendance à ressurgir avec le succès commercial, mais elle reste marginale.

On le retrouve dans toutes sortes de contextes : ateliers autour de tâches partagées tels que ceux qui sont organisés annuellement sous l'acronyme CoNLL, conférence sur l'apprentissage de la langue naturelle ; évaluation ouverte des

systèmes de recherche par mots-clés et de traduction automatique (OpenKWS et OpenMT) ; défi REPERE en France sur la REconnaissance de PERsonnes dans des Emissions audiovisuelles ; *Speaker Recognition Evaluations* et *Text Retrieval Conference* (proposées périodiquement par le NIST) ; *Shared Task on Pronoun Translation* ; *TREC Video Retrieval ; IMAGENET Large Scale Visual Recognition*, et bien d'autres encore. J'en découvre chaque semaine ; il s'agit parfois simplement de partager des bases de données et des critères d'évaluation, comme dans les campagnes TAC (*Text Analysis Conference*) : personne n'est payé pour mener les recherches, celles-ci prennent la forme d'ateliers d'évaluation destinés à encourager la recherche sur le traitement automatique du langage naturel (TALN) grâce à la mise à disposition du public d'un large échantillon de tests, de procédures d'évaluation communes et d'un forum permettant aux divers groupes de partager leurs résultats. TAC recouvre un ensemble de tâches, appelées « tracks », ayant chacune pour objet un sous-problème particulier du TALN. Ces « pistes » explorées dans le cadre de TAC concernent essentiellement des tâches correspondant à des cas d'usage concrets pour un utilisateur, mais elles couvrent également l'évaluation de composants en lien avec ces cas d'usage. Récemment, par exemple, une campagne a été lancée sur la tâche de peuplement de base de connaissances et la synthèse biomédicale. TRECVID porte sur l'analyse et la recherche de contenus vidéos et en pratique couvre l'indexation sémantique, la recherche interactive d'événements en vidéosurveillance, la recherche d'objets, de personnes ou de lieux particuliers dans un sitcom de la BBC, la détection multimédia d'événements, la localisation, le repérage d'hyperliens vidéo.

La base de données de Google Street View contenant des numéros de maisons est un cas récent et très intéressant. Google a pris les numéros reconnus de manière automatique par Google Street View, puis a publié un ensemble de plus de 73.000 chiffres dont l'identification avait été faite par des humains. Pour un autre ensemble de 26.000 chiffres, l'information n'a pas été publiée, à des fins de test. Un dernier ensemble contenant un peu plus d'un demi-million d'exemples est destiné à approfondir l'apprentissage. Les performances ont progressé de manière extraordinairement rapide. En 2011, le taux d'erreur était de 36,7%, il n'était plus que de 1,92% en 2015. Les progrès ne sont pas toujours aussi rapides mais cette méthode garantit en tous cas des progrès réguliers. Voici un échantillon des numéros.

Image 2 : Échantillon d'un ensemble de numéros de maisons pris par Google Street View

La figure suivante représente un célèbre graphique qui retrace l'histoire des tests de performance de la transcription de la parole jusqu'à 2009. Ce qui est remarquable, c'est que dans la plupart des cas, les courbes s'orientent vers le bas. Lors des premiers essais de tâche avec le système de reconnaissance vocale conversationnelle ***Switchboard*** au début des années 1990, le taux d'erreur avoisinait les 90%. Loin de se décourager les chercheurs s'en réjouirent car cela leur assurait de longues années de recherches. En effet, quand on essaie quelque chose de nouveau, si le taux d'erreur est de 10%, sachant que 5% correspond au bruit de fond, cela signifie qu'on n'a pas beaucoup d'années de financement devant soi. Les progrès de la transcription de la parole en texte sont à peu près continus si bien que le corpus de ***Switchboard***, qui a connu une période de stagnation pendant une quinzaine d'années avec un taux d'erreur de 20 à 30%, se situe désormais autour de 10%, ce qui le rapproche de ce qui se passe dans toute conversation humaine : les désaccords sur ce qui a été effectivement dit concernent environ 5 à 6% de tout discours.

Lors de la première conférence sur le traitement appliqué des langues naturelles en 1983, aucune des 34 présentations n'utilisait alors de données publiées ou de système d'évaluation formel. En 2010, lors de la 48ème rencontre annuelle de l'***Association for Computational Linguistics***, les 274 présentations s'appuyaient des données et des méthodes d'évaluation publiées à l'exception de trois d'entre elles pour la bonne raison qu'elles portaient sur la création de nouvelles bases de données ou de nouveaux systèmes d'évaluation.

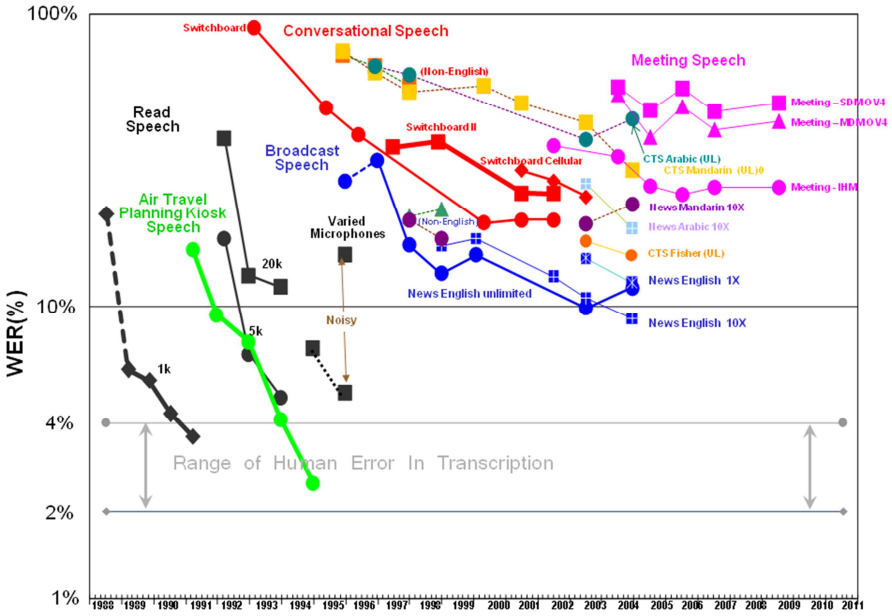


Figure 1 : Historique des tests du NIST en reconnaissance de la parole, mai 2009

J'espère que ce texte aura donné un aperçu historique d'une culture désormais fermement ancrée et bien établie parmi les spécialistes du traitement automatique du langage naturel. Les chercheurs les plus jeunes n'ont pas toujours conscience qu'il a pu en être autrement dans le passé. Qu'en est-il de cette histoire culturelle dans d'autres domaines, du moins dans les domaines scientifiques, en dehors de l'ingénierie ? De nombreux domaines scientifiques ont suivi le même type d'évolution car la mise en commun des données et des problèmes permet de faire baisser les coûts et les barrières à l'entrée. Prenons l'exemple de recherches qui ont beaucoup intéressé les États-Unis comme le reste du monde récemment : les projections concernant le développement de maladies neurodégénératives chez les personnes âgées, en particulier la maladie d'Alzheimer. Imaginons que j'aille chez le médecin et que je lui explique que j'ai de plus en plus de mal à me souvenir des numéros de téléphone et des noms et que je me demande si j'ai quelque chose qui ne va pas. Le médecin me fera passer un test. Mettons qu'il découvre que ma mémoire immédiate des nombres est un peu en dessous de la moyenne, voire assez faible pour

une personne de mon âge. Je voudrai qu'il m'explique ce qui m'attend. Éventuellement, il me fera faire une analyse de sang, il me fera passer une IRM pour voir l'état de mon cerveau, il fera prélever du liquide cérébrospinal, commander le séquençage de mon ADN pour en fin de compte ne rien m'apprendre sinon que d'ici 10 ans, je serai peut-être un légume ou alors je serai exactement comme aujourd'hui, si ce n'est que j'aurai perdu encore un peu de mémoire.

D'un certain point de vue, cela ressemble fort à la reconnaissance automatique de la parole : on dispose potentiellement d'un grand nombre de données concernant la façon dont les personnes parlent et écrivent, de données concernant leurs capacités en termes de mémoire immédiate (même s'il existe bien d'autres tests psychologiques), leur formule sanguine, leur cerveau, leur génome, etc. Ce genre de démarche, à partir de données longitudinales portant sur un grand nombre de personnes, devrait permettre de faire des prévisions. Mais cela ne saurait se faire à moins de disposer d'un groupe clinique comportant un très grand nombre de patients de ce type. Or de tels groupes n'existent pas, les plus importants ne comportent au plus que quelques centaines de patients. Par conséquent, en 2004, le NIST a mis en place une étude, l'*Alzheimer's Disease Neuroimaging Initiative* (ADNI), impliquant 30 sites cliniques[13]. Neil Buckholtz du *National Institute on Aging* fait partie de l'équipe qui a lancé le projet. Lors d'une conférence à Berlin en 2011, nous faisions partie du même comité, « *Transforming Research through Open Access to Discovery Inputs and Outputs* » ; sa communication, qui portait sur l'ADNI, me surprit énormément, notamment lorsqu'il nous présenta certains des buts de l'étude d'observation longitudinale multi-sites de l'ADNI :

- Collecter des données et des échantillons afin d'établir une base de données cliniques sur les IRM cérébrales et les bio-marqueurs en vue d'identifier les marqueurs les plus appropriés pour suivre l'évolution de la maladie et la réaction au traitement.
- Déterminer quelles sont les méthodes optimales pour collecter, traiter et distribuer les images obtenues par IRM ainsi que les bio-marqueurs

[13] http://adni.loni.usc.edu/

conjointement avec des données cliniques et neuropsychologiques dans un contexte de sites multiples.

- « Valider » les données concernant les images obtenues par IRM et les bio-marqueurs par corrélation avec les données neuropsychologiques et cliniques.
- Rendre accessibles au public le plus rapidement possible les échantillons ainsi que la *totalité* des données.

Pour obtenir ces données, il ne suffit pas de les télécharger depuis Internet. Il faut aller sur leur site et leur écrire pour se présenter, expliquer pourquoi on voudrait obtenir leurs données et ce que l'on se propose d'en faire. Pourvu que l'on respecte cette démarche, on obtient les données sans difficulté. Et pour peu que l'on trouve l'idée qui permettra de faire des pronostics meilleurs que ceux des autres sur l'évolution des maladies neurodégénératives, on est en lice pour le Prix Nobel de médecine. Cependant, il est dommage qu'il ne soit pas possible de distinguer entre les différentes versions des bases de données mises en ligne par l'ADNI. Une fois signée l'acceptation des termes et conditions, l'ADNI vous envoie toutes les données accumulées jusqu'au moment de votre signature. Et faute de système d'évaluation, il est impossible de comparer ses propres résultats avec ceux des autres. On vous envoie tout en vrac car il va apparemment de soi que les chercheurs en biologie médicale sachant comment s'y prendre, ils n'ont pas besoin qu'on leur fournisse de système d'évaluation. Il n'existe pas non plus d'atelier dédié dans lequel les chercheurs peuvent comparer leurs résultats avec ceux des autres. L'ADNI ne publie ses résultats que quand elle le juge opportun. À mon avis, prédire l'évolution dans le temps de la maladie d'Alzheimer est exactement le genre de problème pour lequel la méthode de la tâche partagée semble bien marcher.

Il me semble qu'on pourrait envisager d'appliquer de telles méthodes à la classe assez vaste de problèmes similaires dans le domaine biomédical. Au début, les scientifiques seraient sans doute horrifiés ; ils réagiraient peut-être à la manière de Richard Schwartz : ils penseraient que les bureaucrates sont en train de leur ôter leur liberté de chercheurs pour les contraindre tous à suivre une même démarche imposée. Mais il faudrait peut-être tout de même tenter le coup.

La collection *Prismes*

La collection *Prismes* rassemble des textes originaux qui traitent de questions théoriques contemporaines. Ses auteurs sont des contributeurs aux conférences, aux débats ou aux séminaires du Centre et de la Fondation.

Dernières parutions

35. Les big data changent-ils la donne en finance ?
Mathieu Rosenbaum

34. Saturation et croissance : Quand la demande en matières premières stagne
Raimund Bleischwitz et Victor Nechifor

33. Une analyse temps-fréquence des données des prix de pétrole
Josselin Garnier et Knut Sølna

31. *Comment fuir le long d'une droite*
Laure Dumaz

30. Les statistiques peuvent-elles se passer d'une théorie des probabilités ?
Noureddine El Karoui

La liste complète des publications se trouve sur le site
www.centre-cournot.org

Mark Liberman est Professeur de phonétique, de linguistique et d'informatique à l'Université de Pennsylvanie et le fondateur et directeur du *Linguistic Data Consortium*. Il a obtenu sa thèse en linguistique au Massachusetts Institute of Technology (MIT) en 1975 et a commencé sa carrière dans les laboratoires Bell d'AT&T la même année. Il a été membre de leur équipe de 1975 à 1987 et directeur du Département de la recherche en linguistique de 1987 à 1990. Fondateur du carnet en ligne le plus consulté au monde, le *Language Log*, il est l'auteur de trois livres et d'une centaine d'articles.

Ses recherches portent sur la phonétique de type corpus ; le traitement automatique de la parole et du langage ; la phonologie et la phonétique du ton lexical et sa relation à l'intonation ; les moyens gestuels, prosodiques, morphologiques et syntactiques pour marquer l'accent, et leur utilité dans le discours ; les modèles pour l'annotation linguistique ; la récupération et l'extraction de l'information de textes ; l'organisation de la communication parlée dans le cerveau humain, particulièrement au regard des substrats évolutionnistes de la parole et du langage, et par rapport aux systèmes analogues chez d'autres animaux. Mark Liberman s'intéresse également à la modélisation de l'évolution du langage et de l'apprentissage.

Publication sous la direction de Jean-Philippe Touffut
Illustration : Gerald Wassen / Mise en page : Sophie Otrage