

اللسانيات الحاسوبية - الترجمة الآلية [*]

لورانس دانلوس

ترجمة : د. رضا بابا أحمد ، جامعة معسكر ، الجزائر

ملخص

هذه ترجمة لقال للباحثة الفرنسية لورانس دانلوس تقدم فيه عرضا تاريخيا ومنهجيا لللسانيات الحاسوبية والترجمة الآلية بأسلوب بيداغوجي سلس وممتع، وتبث إنجازات الباحثين في هذين المجالين فنتقد ما حققوه وتشير إلى ما ينتظرهم من جهد. وقد استهلكت ذلك بالإشارة إلى التطور الهائل الذي حققته التطبيقات الحاسوبية، واستبعدت إنجاز حاسوب بإمكانه الفهم في الوقت الحالي. وقد مثلت لتلك التطبيقات ببرمجتين هما إليزا وريا حيث تعد الأولى بحق نموذجا يحتذى به كل البرمجيات اللاحقة وتسير بتقنية نمط التطابق، أما الثانية فتتمثل أوج ما بلغته الأبحاث في الذكاء الاصطناعي بحيث يمكنها التماثل مع المستخدم وفق لغة طبيعية ولكن ضمن نطاق خاص. وفي الجزء المخصص للترجمة الآلية، أكدت الباحثة على فشل الترجمة الحرفية للنصوص لأنه لا يمكن أن نترجم بدون فهم، وبالتالي علينا توفير مقياس للتحليل وآخر للتوليد في النظام حتى تكون الترجمة مناسبة، غير أن اللغة تحمل في طياتها لبسا يعرقل هذه العملية وذلك في جميع مستوياتها، وقد مثلت لذلك باللبس الحاصل في المشتراكات الخطية. وفي الأخير، ختمت الباحثة مقالها بالإشارة إلى أن تلك الصعوبات لا تقف حاجزا أمام المختبرات والمؤسسات التي تعمل بسعي حثيث من أجل تزويدنا ببرمجيات متطورة للترجمة الآلية، لكن بلوغ ذلك المرام يحتاج إلى المزيد من مجهودات الباحثين في صورة مختلف المعارف التي تتضمنها اللغة.

الكلمات الجوهرية : المعالجة الآلية للغة، اللبس، المشترك الخطي .

Abstract

The French researcher Laurence Danlos provides us with a historical and methodological overview of computational linguistics and machine translation in a nice pedagogical style. In fact, the researcher tackled the achievements of researchers in these areas in a critical way. She excluded the idea of a computer achievement and understanding at the moment. Indeed, her application where through two programs (named Eliza and Rea). The former is the last and best model for all subsequent software and works with the technology of parallelism. However the latter is the tremendous and highest investigation in artificial intelligence so that they can interact with the

user according to the Natural Language, within a particular domain. Laurence indicates the failure of the literal translation of the texts and excluded the idea that of translating without understanding at the moment. Therefore we need to provide a measure of analysis and another for the generation of the system until the translation is appropriate. However, the language carries the confusion and thus, it hinders this process at all the levels. (i.e. We can notice this in the linear participant). Ultimately, the researcher concluded her essay by pointing out that these difficulties do not stand as a barrier to laboratories and institutions that are working actively. Thus, to reach the goal needs more efforts from researchers in different knowledge that contained a language.

Keywords : Natural language processing, ambiguity, homograph

في الفيلم الشهير « 2001 : أوديسة الفضاء » لس.كوبريك، يفهم الحاسوب هال الإنسان، ويتجاوز معه بلغته، وينقذ تعليماته، ويشعر بأحاسيسه. في سنة 1968⁽¹⁾، اعتقد مارفين منسكي، وهو المستشار العلمي للفيلم بصفته متخصصا في الذكاء الاصطناعي، بأنه بالإمكان إنجاز حاسوب حقيقي مثل هال في سنة 2001. فهل نقرب ونحسن في سنة 2000⁽²⁾ من إنجاز هال؟ نحاول في هذا المقال أن نبدي ونوضح أسباب بعدنا عن هذا الإنجاز؟

يبدو على الفور بأن هذا الإقرار يمكن انتقاده في الوقت الذي تكتسح فيه تطبيقات اللسانيات الحاسوبية (أو الذكاء الاصطناعي على العموم) بإيقاع متسارع حياتنا اليومية. في الواقع، هناك عدد لا بأس به من السيارات المزودة بجهاز ناطق، وكل معالج نص يتضمن مدققا إملائيا، وأنظمة الإملاء والترجمة الآليين مزدهرة في السوق، ونتوقع بانتظام صناعة آلات تصوير وهواتف مترجمة ومحركات بحث في الشبكية تقترح ملخصا أو ترجمة للنصوص المطلوبة، كما يمكن أن نستفيد من مساعدة شخصية افتراضية من أجل تدبير موعد ما، إلخ. باختصار: في الواقع، توجد اليوم تطبيقات لم يكن بالإمكان تخيلها قبل أربع أو خمس سنوات. لماذا نحن إذن بعيدون عن هال؟ الإجابة عن هذا السؤال تستند على مفهوم رئيسي هو: الفهم. يتطلب إنجاز هال أن يفهم الحاسوب اللغة، في حين أن التطبيقات المقترحة اليوم لا تفترض فهما أو تفترض فهما سطحيًا جدا. وههنا يكمن الفرق إذ ليس بالإمكان في الوقت الراهن محاكاة عملية فهم اللغة في الحاسوب.

سُقسّم هذا المقال إلى جزأين. في الجزء الأول، سنقدم تطبيقات يتعامل فيها الحاسوب مع اللغة المنطوقة أو المكتوبة دون فهم أو بفهم سطحي. سنقدم انطلاقا من ههنا ما يمكن أن يكون عليه الفهم بالنسبة للحاسوب وسنتحدث عن «العمر الذهني للإنسان الآلي».

في الجزء الثاني، سنقدم تطبيقاً خاصاً، هو الترجمة الآلية، يحتاج فيه الحاسوب إلى الفهم إذ لا يمكن أن نترجم دون فهم. من خلال هذا التطبيق، سنعرض بعض الصوبات الكبيرة المصادفة في المعالجة الآلية للغة المكتوبة. تفسّر هذه الصعوبات تكرار الأخطاء الملاحظة في النصوص المترجمة آلياً، وتوضّح العمل المنجز والعمل الذي يُنتظر من الباحثين إنجازه لمحاكاة الفهم البشري في الحاسوب.

1. تطبيقات بدون فهم أو بفهم سطحي :

سنقدم تطبيقين أحدهما أنجز سنة 1964 والآخر في أيامنا هذه، سيسمح لنا ذلك بتقدير العمل المنجز والعمل المنتظر من الباحثين.

منذ سنة 1964، أنجز ج. فايزمباوم وزملاؤه في معهد ماستشوستس للتكنولوجيا برمجة إليزا، التي أثّرت ولا تزال مؤثرة حتى اليوم. هذه البرمجية تقترح محاكاة لحوارات مكتوبة تدور بين معالج نفسي وبين مريض، ودور المعالج النفسي يؤديه الحاسوب. فلنقدّم فوراً مثلاً لحوار: إجابات المريض يرقنها شخص على الحاسوب وتكون بـخط مائل، أما إجابات المعالج النفسي فتكون بخط عريض.

- مرحباً، أنا معالجك النفسي. حدثني عن مشاكلك.

- أنا قلق جداً.

- لماذا تقول بأنك قلق جداً ؟

- بسبب زوجتي.

- حدثني عن عائلتك.

- أحسب أولادي.

- أرغب في أن تحدثني عن علاقتك مع أولادك.

- لا يحببني أحد.

- هل يحزنك أن لا يحبك أحد ؟

- نعم، أحس بالوحدة.

- هل لحياتك الجنسية دخل في ذلك ؟ واصل بكلّ حرية.

...

علّقت على هذه البرمجية آمال، فالحوار مدهش، مع أن الحاسوب المعالج النفسي لا يفهم من إجابات مريضه شيئاً على الإطلاق. كيف تمّ ذلك الحوار إذن؟ عبر تقنية نمط التطابق التي تتضمنها مبادلات الحوار والتي سنشرحها باختصار.

« مرحباً، أنا معالجك النفسي. حدثني عن مشاكلك. » هذا المقطع هو أحد النصوص المسجلة سلفاً الذي يستهلّ به الحاسوب المعالج النفسي حوارهِ. يُختار هذا النص عشوائياً من مجموعة نصوص مسجلة ذات مضمون متماثل.

« أنا قلق جداً. » هذا المقطع هو إجابة حرة للمريض تُخزّن في متغير R1.

« لماذا تقول بأنك قلق جدا؟ » هذا السؤال من المعالج النفسي ليس إلا تسوية تركيبية صرفية لسلسلة الرموز التالية: لماذا تقول R1 ؟ الحاسوب ليس بحاجة لأن يفهم معنى R1 لوضع هذا السؤال، وإنما يكتفي بتحويل « أنا قلق جدا » إلى « أنك قلق جدا » ؛ إذ لو قال المريض: « أنا مسرور جدا » لاستفهم منه « المعالج النفسي » ببساطة قائلا: « لماذا تقول بأنك مسرور جدا؟ » بسبب زوجتي « هذا المقطع هو إجابة حرة للمريض يتضمن المفردة «زوجة».

« حدثني عن عائلتك » هذا المقطع يأتي في ترتيب نظامي لدى المعالج النفسي حالما يتضمن مقطع المريض إحدى مفردات القائمة: «زوجة»، «زوج»، «أب»، «أم»...
 « نعم، أحس بالوحدة » هذا المقطع هو إجابة حرة للمريض لا تتضمن شيئا مهماً.
 « هل لحياتك الجنسية دخل في ذلك؟ واصل بكل حرية. » هذا رد المعالج النفسي عندما لا يدري ما يقول.

مجمل القول : إن الحاسوب المعالج النفسي لا يفهم شيئا من مداخلات مريضه، بل يكتفي بإنتاج إجابات عبر تفعيل إجابة من بين مئات أو آلاف الإجابات المسجلة وعبر إقامة تحويلات تركيبية صرفية (على العموم: « أنا » تتبعها « أنت »).

قريبا من 40 سنة بعد إليزا، يعمل حاليا ج.كاسل وزملاؤه في معهد ماستشوستس للتكنولوجيا على إنجاز برمجية رنا، التي تمثل أوج البحث والتكنولوجيا. هذه البرمجية تقترح محاكاة لحوارات شفهية بين سمسار عقارات وبين عميله. يؤدي دور السمسار إنسان آلي، أي آلة في شكل إنسان قادر على أن يتواصل بالإيماءات والنظرة والكلام (ما يعرف بمصطلح «الاتصالات متعددة الوسائط»). التقدم كبير ما بين إليزا (1964) ورنا (2000)؛ فلقد حدث انتقال من حوار مكتوب إلى حوار شفهي (نظرا للتقدم الهائل في التعرف على الكلام واصطناعه)، كما حدث انتقال من حاسوب إلى إنسان آلي (نظرا للتقدم الهائل في علم الإنسانيات الآلية)، وأخيرا حدث انتقال إلى اتصال متعدد الوسائط بفضل إدماج جيد لمختلف التكنولوجيات. لكن ماذا عن الفهم ؟ التقدم ضعيف بالنسبة للفهم، رنا تفهم تقريبا ما يتحدث عنه عميلها، لكن إذا اقتصر على المسائل التي تتعلق بالعقار فقط. فإذا انتقل العميل من شراء شقة إلى شراء سيارة، فإن رنا تتوه بالكلية. يرجع ذلك إلى أن مقياس الفهم لا يتضمن إلا معارف لسانية وعقلية حول العقار، وينحصر معجمه في مفردات العقار خاصة. بصفة أكثر تعميما، ليس هناك أي نظام عام للفهم، بمعنى أنه ليس هناك نظام يشتغل من أجل الحديث المتداول وفي عدة نطاقات محددة (العقاري، القانوني، الطبي...). في الوقت الراهن، لا يستطيع الحاسوب أن يفهم (أي أن يحسب تمثيلا دلاليا لهذا النص مجردا بما فيه الكفاية حتى يمكن له أن يقوم بتفكير حوله) إلا إذا انحصرت مفردات ذلك الحديث إلى نطاق محدد، وفي هذه الحال يكون محصورا من الناحيتين اللسانية والمفهومية. ستوضح الأسباب التقنية لهذه التحديدات في الجزء الثاني من المقال.

غالبا ما يُعبر في وسائل الإعلام عن هذه التحديدات لفهم الإنسان الآلي بالقول: « عمر رنا الذهني ثلاث سنوات. » نحن نعتقد بأن هذا النوع من الإقرارات خاطئ في الأساس؛ لا لأن

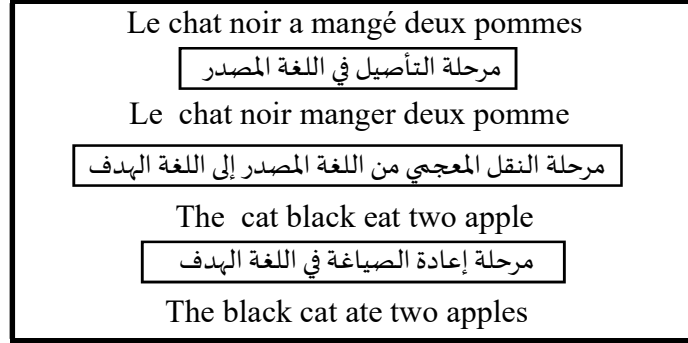
العمر الذهني هو سنتان أو أربع عوضاً عن ثلاث، ولكن ببساطة لأن من الخداع عقدَ مقارنة بين العمر الذهني للإنسان الآلي وبين العمر الذهني للطفل. في الواقع، ليس هناك فيما نعلم أي طفل لا يستطيع فهم اللغة إلا ضمن نطاق محصور (على العموم التغذية) ويقصي جميع النطاقات الأخرى (على العموم الألعاب، العناق...) بمن في ذلك من هو مصاب باضطراب لغوي. بعبارة أخرى: لا يمكن الحديث عن العمر الذهني للإنسان الآلي مقارنة بالعمر الذهني للطفل. تعلّم الطفل للغة (وإدراكه للعالم) يمرّ بآليات لا تزال غير معروفة جيداً حتى الآن لكن ليس لها بالضرورة علاقة بالآليات المستعملة في إنجاز الإنسان الآلي مثل ريا التي خُصرت معرفتها عن قصد في نطاق العقار دون محاولة البحث على الأقل عن توسيع حقل كفاءاتها. من الشائع أيضاً الحديث عن «الحالات المزاجية» للإنسان الآلي. وهكذا «يُوح لك الإنسان الآلي بحالاته المزاجية» -كما يدّعي- عندما يقول: «نفذت بطارياتي». ولكن بعد ذلك سيارتك أيضاً تبوح لك بحالاتها المزاجية عندما يشتعل الضوء الأحمر للبطارية! المبدأ هو نفسه لا تختلف إلا طريقة التواصل وصورة الموضوع. هل يكفي أن يكون الموضوع ذا صورة إنسانية ويصطنع النطق حتى تكون له حالات مزاجية؟

2. الترجمة الآلية :

في نظام للترجمة الآلية، يقدّم نص باللغة المصدر (الفرنسية مثلاً) في مدخل النظام في شكله الإلكتروني، فيقوم النظام بحساب النص باللغة الهدف (الإنجليزية مثلاً). هذه الترجمة التي تظهر في شكل إلكتروني يمكن أن تُطبع أو تُنشر في الشبكة.

الترجمة الآلية أقدم تطبيق للسانيات الحاسوبية — حيث عاصرت البحوث في هذا المجال بدايات المعلوماتيات نفسها (في أواخر سنوات 1940). هذا التطبيق دقيق جداً لأن الاحتياجات إليه كثيرة، فالمفوضية الأوروبية وحدها تترجم حوالي مليون صفحة سنوياً، وباقي الجنسيات حوالي المليار صفحة.

اعتمدت الأنظمة الأولى للترجمة الآلية نقلاً حرفياً ممثلاً في الشكل 1. تتضمن مرحلة التأصيل في اللغة المصدر رد الأسماء التي في حالة الجمع إلى المفرد، ورد الأفعال المضارعة إلى ماضيها... وتتضمن مرحلة النقل المعجمي صلة الكلمة في اللغة المصدر بمقابلتها في اللغة الهدف. وتسمح مرحلة إعادة الصياغة في اللغة الهدف باحترام قواعدها التركيبية والصرفية (على العموم النعت يسبق المنعوت في الإنجليزية).

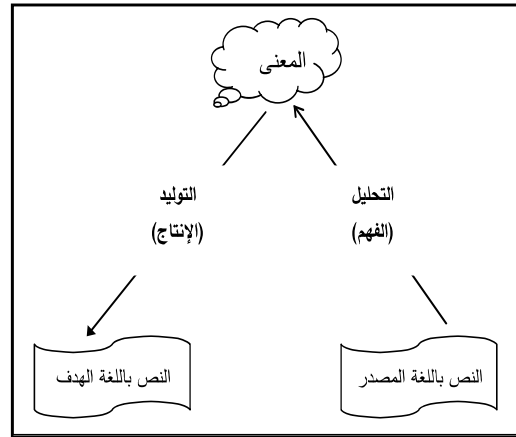


الشكل 1: الترجمة الحرفية

نعرف جيدا حدود الترجمة الحرفية التي تعطي للجملة (1) أدناه ترجمة خاطئة (2) عوضا عن الترجمة الصحيحة (3).

- (1) Un pied-noir a mangé une pomme de terre. (أكل المعمر بطاطس).
 (2) A black foot ate an apple of earth.
 (3) An Algerian-born Frenchman ate a potato.

إن الترجمة الحرفية مآلها الفشل لأن من المعروف عدم قدرتنا على الترجمة دون فهم. لذلك يجب أن يتضمن نظام الترجمة الآلية (مبدئيا) مقياسا للتحليل ومقياسا للتوليد (الشكل 2). يتولى مقياس التحليل فهم النص باللغة المصدر وحساب تمثيله الدلالي، ثم يُنقل هذا التمثيل الدلالي إلى مدخل مقياس التوليد الذي يقوم بإنتاج النص باللغة الهدف.



الشكل 2: معمارية نظام الترجمة الآلية

إن هذه المعمارية ترتبـط بالمبادئ أكثر من ارتباطها بالواقع لأننا لم نصل إلى إنجاز مقاييس للتحليل والتوليد إلا جزئياً. لماذا؟ لأن اللغة تعكس لبساً لا ينتهي. العلاقة بين الشكل والمعنى ليست تقابلية تناظرية؛ فللشكل اللساني عدة معانٍ من جهة (مما يورث التباساً في التحليل)، ومن جهة أخرى يمكن التعبير عن المعنى بعدة ألفاظ (مما يورث التباساً في التوليد). سنمثّل لللبس اللساني في التحليل فقط وفي حالة الاشتراك الخطي، أي الغموض النابع عن كتابة كلمتين لهما الشكل الإملائي نفسه لكنّ معنييهما مختلفان (وبالتالي عموماً لهما ترجمتان مختلفتان). وعليه، فإن « le » إما محدّد وإما ضمير، و « savoir » إما اسم وإما فعل، أما « avocat » فدائماً اسم ولكن إما أن يدل على المحامي (فيترجم إلى lawyer بالإنجليزية) وإما أن يدل على نوع من الفواكه (فيترجم إلى avocado بالإنجليزية)⁽³⁾. في معظم الحالات، لا يثير معنياً « avocat » لبساً حقيقياً: في الجملتين (4) و (5) أدناه، يكشف الإنسان بسهولة الالتباس في هذه اللفظة بفضل السياق الذي استُخدمت فيه، ويثبت أنها تدل على نوع من الفاكهة في (4) وعلى المحامي في (5).

(4) Zoé a mangé un avocat. (أكل زوي فاكهة الأفوكاتو)

(5) Zoé a rendez-vous avec un avocat. (لزوِي موعدٌ مع محامٍ)

لكنّ هناك لبساً افتراضياً لدى الحاسوب يُسهّم فيه مقياس التحليل الذي يتضمن معني « avocat ». يجب حتماً أن يُرفع هذا اللبس الافتراضي حتى نتجنّب ترجمات خاطئة مثل (6) و (7).

(6) Zoe ate a lawyer. (أكل زوي محامياً)

(7) Zoe has a meeting with an avocado. (لزوِي موعدٌ مع فاكهة الأفوكاتو)

رفع اللبس الافتراضي في المشتراكات الخطية يتطلب القيام بمهمتين هما:

1 - إقامة تصنيف دلالي لأسماء الأعلام (تكتب الأسماء بحروف مكثّرة)، مثلاً: avocat ينتهي إلى صنف FRUIT الذي بدوره ينتهي إلى COMBUSTIBLE (مأكولات) و avocat ينتهي إلى صنف HUMAIN (إنسان)؛

2 - تحديد المقولة الدلالية لمكثّلات الفعل مثلاً :

ANIME manger COMBUSTIBLE

HUMAIN avoir rendez-vous avec HUMAIN⁽⁴⁾

لكن المهمة معقدة لأن الأفعال غالباً ما تكون هي أيضاً مشتراكات خطية، وعليه فإن manger له معانٍ مختلفة الممثل منها في (4) وفي الأمثلة التالية :

(8) Ce poêle mange beaucoup de charbon. (هذه المدفئة تحرق فحماً كثيراً)

APPAREIL manger COMBUSTIBLE

This stove uses a lot of coal.

(9) Les grosses entreprises mangent les petites. (المؤسسات الكبرى تبتلع الصغرى)

ORGANISATION manger ORGANISATION

Big firms swallow up smaller ones.

لذلك فإن الجملة (4) تتضمن التباسا مركبا إذ لا يمكن رفع لبس avocat حتى يرفع لبس manger والذي يتوقف بدوره على رفع لبس avocat. هذه الالتباسات المركبة تولّد عددا كبيرا من التوليفات يمكن أن نرقمها كالآتي :
إذا احتوت جملة ما على n كلمة

$$m_1 m_2 \dots m_i \dots m_n$$

وإذا كان لكلمة $m_i k_i$ من المعاني (وبالتالي عموما k_i من الترجمات)، فإنه ينبغي على الحاسوب أن يختار من بين K فرضية بحيث :

$$K = k_1 \times k_2 \times \dots \times k_i \times \dots \times k_n$$

نقول: إن مقياس التحليل يولّد بإفراط وينتج عددا هائلا من الفرضيات. في بعض الأحيان، لا يمكن رفع اللبس عن المشترك الخطي بالاعتماد على السياق القريب (الجملة التي يظهر فيها) ولا حتى بالاعتماد على سياق أوسع. ومنه في (10) لا يمكن أن نرفع اللبس الموجود في « aimer » (أحب) أو في « véreux » (دائد أو فيه الدود) لأن هذين المحمولين ينطبقان على الفواكه والإنسان معا:

(10) Zoé a aimé cet avocat. Pourtant il était véreux.

(أحب زوي هذا المحامي أو فاكهة الأفوكاتو على رغم أنه كان دائدا أو فيه الدود)

Zoé loved/liked this lawyer/avocado. However, he/it was shady/worm-eaten.

لنشدّ على ما يلي: الجملة في (10) غامضة في الحقيقة، لكنها لا تُلمح كذلك في وضعية التلفظ إلا نادرا، مثلا في حوار بين شخصين لأنهما يدركان ما يتحدثان عنه أهو عن حب زوي أو عما أكله في الغذاء. في الترجمة الآلية أو في الفهم عموما، يمكن محاكاة هذا النوع من المعارف بتحديد نطاق النصوص المعالجة: النظام الذي يحتوي على مقياس للتحليل لا يمكنه إعطاء نتائج مرضية إلا في نطاق محدد كالقانون أو المعلومات أو العقار. في الواقع، يمكن لمقياس التحليل المخصص للقانون ألا يحفظ إلا معنى واحدا لـ avocat هو «رجل القانون»، وبالتالي يقلل من عدد الفرضيات الممكنة التي يثيرها الاشتراك الخطي. هذا المنهج الذي تُستقرأ فيه بعض الأخطاء هو نقطة عبور ضرورية لأن المشتركات الخطية ليست إلا مثالا للبس الموجود في اللغة، فلنتذكر ذلك، هو لبس من النوع الدلالي لكن توجد أنواع أخرى من اللبس الدلالي (على العموم أزمنة الأفعال كاستعمال فعل في الزمن الحاضر لواقعة مستقبلية في « Zoé vient demain » (يأتي زوي غدًا)⁽⁵⁾ وأنواع من اللبس في جميع مستويات اللغة الصرفي منها والتركيب والتداولي. هذا السيل من اللبس الذي يغمر النص من خلال شبكة معقدة من اللبس المتداخل الذي لا تمكن إدارته غير محاكاة الفهم البشري القادر على ضبط المحادثات العادية والمحادثات الأكثر تقنية بالقدر نفسه.

رغم ذلك، فإن أنظمة الترجمة الآلية تتطور شيئاً فشيئاً بخطوات ثابتة داخل تيارين. في مختبرات البحث، طوّرت عدة طُرُز تعطي نتائج نوعاً ما مرضية في نطاقات ضيقة باستخدام حواسيب ذات قدرة وسرعة فائقتين. وفي الأسواق تنتعش منتجاتها التجارية والتي على رغم أنها تقدّم ترجمات خاطئة يمكن أن تسهم على الأقل في تحديد ما يتحدث عنه النص (الذي هو مهم في المراقبة التكنولوجية مثلاً). تغطي هذه الأنظمة معاجم واسعة للمفردات تشتغل في الحاسوب الشخصي وتنتج الترجمة فوراً. نعتقد بأن هذين التيارين يمكن أن يتقاربا مع تطور قدرة الحاسوب. بالتأكيد، لكن هذا غير كاف. ينتظر الباحثين إنجاز عمل جبار يتضمن صورنة كل المعارف اللسانية والعقلية⁽⁶⁾ وتخزينها في الحاسوب، ولن ينتهوا من ذلك في 2001.

هوامش الترجمة :

(*) هذه ترجمة للمقال الآتي :

Danlos (Laurence) 2002, Linguistique informatique – traduction automatique, Yves Michaud (eds.), Le cerveau, le langage, le sens, Ed. Odile Jacob, Paris, vol. 5, pp. 261-270.

- (1) تاريخ إصدار هذا الفيلم . المترجم .
- (2) تاريخ كتابة هذا المقال . المترجم .
- (3) في اللغة العربية توجد هذه الظاهرة أيضاً مثلاً: « أجل » - بدون تشكيل - إما اسم (أَجَلٌ)، أو فعل (أَجَلْتُ) أو حرف (أَجَلْ) . المترجم .
- (4) معنى العبارة الأولى أنه حتى يتمّ الأكل manger يجب أن ينتهي الفاعل إلى صنف ANIME (حركي) والمفعول إلى صنف COMESTIBLE، ومعنى العبارة الثانية أنه حتى يؤخذ موعد avoir rendez-vous يجب أن ينتهي الفاعل والمضاف إلى صنف HUMAIN . المترجم .
- (5) هذا النوع من اللبس موجود في العربية أيضاً حيث يدلُّ الفعل المضارع على الحاضر والمستقبل، ولا يخلُص لأحدهما إلا بقرائن . المترجم .
- (6) المعارف اللسانية هي التي يتوصل إليها بالوضع اللغوي، أما المعارف العقلية فبالعقل.

مراجع الترجمة :

1. إدريس (سهيل) 2003، المنهل، دار الآداب، بيروت، ط. 31.
2. معجم الحاسبات 1995، منشورات مجمع اللغة العربية، القاهرة، ط. 2.
3. معجم فيليبس (إنجليزي-عربي) لعلوم وهندسة الحاسوب.
4. <http://www.phillips-publishing.com/generalsupport.html>
5. المعجم الموحد لمصطلحات اللسانيات (إنجليزي-فرنسي-عربي) 1989، مطبعة المنظمة العربية للتربية والثقافة والعلوم، تونس.

قائمة المصطلحات :

بالفرنسية	بالعربية
communications multimodales	الاتصالات متعددة الوسائط
homographie	اشتراك خطي
synthèse de la parole	اصطناع الكلام
robot humanoïde	الإنسان الآلي
Robotique	الإنسانيات الآلية
lemmatisation	التأصيل
Traduction Automatique	الترجمة الآلية
Traduction mot-à-mot	ترجمة حرفية
biunivoque	تقابلية تناظرية
réajustement	تسوية
reconnaissance de la parole	التعرف على الكلام
Intelligence Artificielle	الذكاء الاصطناعي
Prototype	طراز
Ambiguïté	اللبس
Linguistique Informatique	اللسانيات الحاسوبية
langue source – cible	اللغة المصدر - الهدف
Simulation	محاكاة
correcteur orthographique	مدقق إملائي
psychothérapeute	معالج نفسي
Traitement automatique de langue	المعالجة الآلية للغة
MIT	معهد ماستشوستس للتكنولوجيا
connaissance extralinguistique – linguistique	معرفة عقلية (خارج لسانية) – لسانية
module d'analyse – de génération	مقياس التحليل - التوليد
transfert lexical	النقل المعجمي
pattern-matching	نمط التطابق

Linguistique informatique – traduction automatique
par Laurence Danlos

Danlos (Laurence) 2002, Linguistique informatique – traduction automatique, Yves Michaud (eds.), Le cerveau, le langage, le sens, Ed. Odile .Jacob, Paris, vol. 5, pp. 261-270

Dans le célèbre film 2001, l'Odyssée de l'espace de S. Kubrik, l'ordinateur HAL comprend l'homme, dialogue avec lui dans sa langue, exécute ses commandes, et ressent des émotions. En 1968, Marvin Minsky, conseiller scientifique du film en tant que spécialiste de l'intelligence artificielle, pensait qu'on pourrait effectivement réaliser un ordinateur tel que HAL en 2001. En l'an 2000, est-on près de la réalisation de HAL ? Cet article essaiera de montrer et d'expliquer pourquoi on en est loin.

Cette affirmation peut paraître d'emblée critiquable à l'heure où les applications de la linguistique informatique (ou plus généralement de l'intelligence artificielle) envahissent à un rythme accéléré notre quotidien. En effet, bon nombre de voitures parlent, tout traitement de textes intègre un correcteur orthographique, des systèmes de dictée automatique ou de traduction automatique fleurissent sur le marché, on nous annonce régulièrement des photocopieurs ou téléphones qui traduisent, les moteurs de recherche sur le web proposent un résumé et une traduction des textes trouvés, on peut bénéficier d'une assistante virtuelle personnelle pour gérer ses rendez-vous, etc. Bref, des applications que l'on n'aurait même pas imaginées il y a quatre ou cinq ans existent bel et bien aujourd'hui. Pourquoi est-on donc si loin de HAL ? La réponse à cette question repose sur une notion clef : la compréhension. La réalisation de HAL demande que l'ordinateur comprenne le langage, tandis que les applications que l'on propose aujourd'hui ne supposent pas de compréhension ou alors une compréhension très superficielle. Et c'est là où se situe la différence : on ne sait pas simuler à l'heure actuelle la compréhension du langage sur ordinateur.

Cet article sera divisé en deux parties. Dans une première

partie, nous présenterons des applications où l'ordinateur manipule du langage écrit ou oral sans rien y comprendre ou pas grand-chose. Nous présenterons par là même ce que peut être la compréhension pour un ordinateur et discuterons de « l'âge mental des robots humanoïdes ». Dans la seconde partie, nous présenterons une application particulière, la traduction automatique, où l'ordinateur aurait besoin de comprendre : on ne traduit pas sans comprendre. À travers cette application, nous exposerons quelques grandes difficultés rencontrées dans le traitement automatique du langage écrit. Ces difficultés expliquent les erreurs fréquemment observées dans les textes traduits automatiquement et illustrent la distance à parcourir pour simuler la compréhension humaine sur ordinateur.

*Applications sans compréhension
ou avec une compréhension superficielle*

Nous allons présenter deux applications, l'une réalisée en 1964, l'autre de nos jours. Cela nous permettra de mesurer la distance parcourue et la distance qui reste à parcourir.

Dès 1964, J. Weizenbaum et ses collègues de MIT ont réalisé un logiciel, Eliza, qui a impressionné et impressionne encore. Ce logiciel propose une simulation de dialogues écrits entre un psychothérapeute et un patient, le rôle du psychothérapeute étant tenu par l'ordinateur. Donnons tout de suite un exemple de dialogue : les répliques du patient saisies sur l'ordinateur par un être humain sont en italiques, celles du psychothérapeute en gras.

— **Bonjour, je suis votre psychothérapeute. Parlez-moi de vos problèmes.**

— Je suis très angoissé

— **Pourquoi dites-vous que vous êtes très angoissé ?**

— C'est à cause de ma femme.

— **Parlez-moi de votre famille.**

— J'adore mes enfants.

— **J'aimerais que vous me parliez de vos relations avec vos enfants.**

— Personne ne m'aime.

— Est-ce que ça vous ennuie que personne ne vous aime ?

— Oui, je me sens seul.

— **Votre vie sexuelle y est peut-être pour quelque chose. Continuez sans crainte.**

— ...

On s'y croirait ! Le dialogue est époussouflant ! Et pourtant l'ordinateur psychothérapeute ne comprend rien, absolument rien aux répliques de son patient. Comment cela est-il possible ? Par la technique du pattern-matching que nous allons expliquer brièvement sur les échanges de notre dialogue.

« Bonjour, je suis votre psychothérapeute. Parlez-moi de vos problèmes. »

C'est un des textes préenregistrés par lequel l'ordinateur psychothérapeute entame son dialogue. Ce texte est choisi aléatoirement parmi un ensemble de textes préenregistrés au contenu similaire.

« Je suis très angoissé. »

C'est une réponse libre du patient qui est stockée dans une variable **R1**.

« Pourquoi dites-vous que vous êtes très angoissé ? » Cette question du psychothérapeute n'est qu'un ajustement morphosyntaxique de la chaîne de caractères : pourquoi dites-vous que Ri ? L'ordinateur n'a pas besoin de comprendre le sens de **R1** pour poser cette question. Il se contente de transformer « je suis » en « vous êtes ». Si le patient avait dit « je suis très euphorique », le « psychothérapeute » aurait tout aussi imperturbablement demandé : « Pourquoi dites-vous que vous êtes très euphorique ? »

« C'est à cause de ma femme. »

C'est une réponse libre du patient qui contient le mot « femme ».

« Parlez-moi de votre famille.

C'est un ordre systématique du psychothérapeute dès que la réplique du patient contient un mot de la liste : « femme », « mari », « père », « mère », etc.

« Oui, je me sens seul. »

Réponse libre du patient qui ne contient rien de spécial.

« Votre vie sexuelle y est peut-être pour quelque chose. Continuez sans crainte. »

Réponse du psychothérapeute quand il ne sait plus trop quoi dire.

En résumé, l'ordinateur psychothérapeute ne comprend rien aux interventions de son patient : il se contente de produire des réponses en activant une des centaines ou milliers de réponses préenregistrées et en effectuant des transformations morphosyntaxiques (en général « je suis » implique « vous êtes »).

Près de quarante ans après Eliza, J Cassel et ses collègues de MIT sont en train de réaliser un logiciel, Rea, qui est à la pointe de la recherche et de la technologie. Ce logiciel propose une simulation de dialogues oraux entre un agent immobilier et un client. Le rôle de l'agent immobilier est tenu par un robot humanoïde, c'est-à-dire un robot à forme humaine capable de communiquer par gestes, par le regard et par la parole (ce qui est connu sous le terme de « communication multimodale »). Les progrès entre Eliza (1964) et Rea (2000) sont considérables : on est passé d'un dialogue écrit à un dialogue oral (ce qui est dû aux progrès énormes de la reconnaissance et de la synthèse de la parole), on est passé d'un ordinateur à un humanoïde (ce qui est dû aux progrès énormes de la robotique), et enfin on est passé à une communication multimodale grâce à une bonne intégration de différentes technologies. Qu'en est-il de la compréhension ? Les progrès sont minces Rea comprend à peu près ce dont parle son client, mais uniquement si celui-ci se cantonne à des questions concernant l'immobilier. Si le client passe de l'achat d'un appartement à l'achat d'une voiture, Rea est complètement perdue. Ceci est dû au fait que le module de compréhension n'a que des connaissances linguistiques ou extralinguistiques sur l'immobilier. En particulier, son vocabulaire se limite à celui de l'immobilier. D'une manière plus générale, il n'existe aucun système de compréhension générique, c'est-à-dire fonctionnant pour la conversation courante et pour les multiples domaines fermés (immobilier, juridique, médecine, etc.). À l'heure actuelle, l'ordinateur ne peut comprendre un texte (calculer une représentation sémantique de ce texte suffisamment abstraite pour pouvoir effectuer du raisonnement dessus) que si celui-ci relève d'un domaine fermé et donc restreint linguistiquement et conceptuel-

lement. Les raisons techniques de ces limitations seront expliquées dans la seconde partie de l'article.

Ces limitations dans la compréhension des humanoïdes sont souvent traduites dans les médias dans les termes suivants : « Rea a un âge mental de trois ans. » Nous pensons que ce type d'affirmation est fondamentalement erroné. Non pas du fait que l'âge mental est de deux ou quatre ans au lieu de trois, mais simplement du fait qu'il est spécieux de faire une comparaison entre l'âge mental d'un robot humanoïde et celui d'un enfant. En effet, il n'existe à notre connaissance aucun enfant, même atteint de troubles langagiers, qui ne soit capable de comprendre le langage que dans un domaine restreint (en général la nourriture) à l'exclusion de tout autre domaine (en général les jouets, les câlins, etc.). Autrement dit, on ne peut pas parler de l'âge mental d'un humanoïde en effectuant une comparaison avec l'âge mental d'un enfant. L'apprentissage du langage (et de la perception du monde) par un enfant passe par des mécanismes encore mal identifiés à l'heure actuelle mais qui n'ont forcément rien à voir avec les mécanismes utilisés dans la réalisation d'un humanoïde comme Rea dont on restreint sciemment la connaissance à l'immobilier sans chercher le moins du monde à élargir le champ de ses compétences.

Il est aussi fréquent d'entendre parler des « états d'âme » des robots humanoïdes. Ainsi, un humanoïde « vous confie ses états d'âme », raconte-t-on, lorsqu'il dit : « Mes batteries sont à plat. » Mais alors votre voiture vous confie aussi ses états d'âme lorsque s'allume la lumière rouge de la batterie. Le principe est le même, seuls le mode de communication et la forme de l'objet diffèrent. Suffit-il d'être un objet à forme humaine et à synthèse vocale pour avoir des états d'âme ?

La traduction automatique:

Dans un système de traduction automatique (désormais TA), un texte en langue source (notée Ls, le français par exemple) est donné en entrée sous forme électronique au système de TA qui calcule le texte en langue cible (notée Le, l'anglais par exemple). Cette traduction, qui se présente sous forme électronique, est prête à être imprimée ou diffusée sur le web.

La TA est l'application de linguistique informatique la plus ancienne

- les recherches en TA sont contemporaines des débuts mêmes de l'informatique (vers la fin des années 1940). C'est une application très prisée car les besoins sont énormes. Ainsi la Commission européenne traduit environ un million de pages par an, les multinationales environ un milliard.

Les premiers systèmes de TA reposaient sur une traduction mot à mot qui est schématisée dans la figure 1. La phase de « lemmatisation » en Ls consiste à mettre les noms au singulier, les verbes à l'infinitif, etc. La phase de transfert consiste à associer à un mot en Ls sa traduction en Lc (en général « noir » -s black). La phase de réajustement en Lc permet de respecter les règles morphosyntaxiques de Lc (en général un adjectif se place avant le nom en anglais).

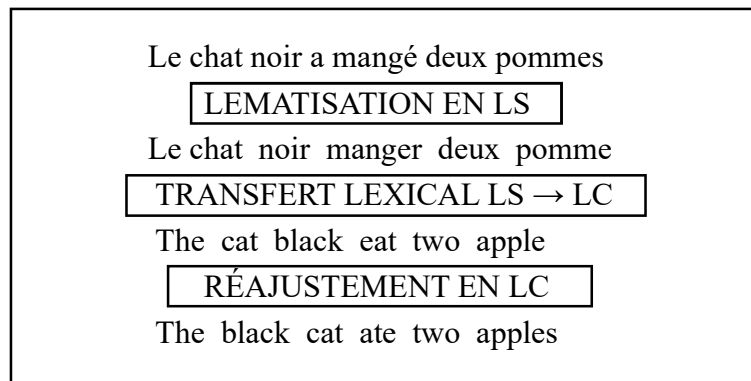


Figure 1 - Traduction mot à mot.

On ne connaît que trop bien les limites de la traduction mot à mot qui donne pour la phrase (1) ci-dessous la traduction erronée (2) au lieu de la traduction correcte (3).

- (1) Un pied-noir a mangé une pomme de terre.
- (2) *A black foot ate an apple of earth.*
- (3) *An Algerian-born Frenchman ate a potato.*

La traduction mot à mot est vouée à l'échec car il est connu que l'on ne peut pas traduire sans comprendre. Un système de TA

doit donc (en principe) comporter un module d'analyse et un module de génération (Fig. 2). Le module d'analyse est chargé de la compréhension du texte en Ls et calcule une représentation sémantique de ce texte. Celle-ci est fournie en entrée au module de génération qui produit le texte en Le.

Mais cette architecture d'un système de TA relève plus du principe que de la réalité car on n'arrive guère à réaliser des modules d'analyse ou de génération. Pourquoi ? Parce que la langue comporte une infinité d'ambiguïtés. La relation sens-forme n'a rien de biunivoque : d'une part, une forme linguistique donnée a plusieurs sens (ce qui crée des ambiguïtés en analyse), d'autre part un sens donné peut s'exprimer par plusieurs s (ce qui crée des ambiguïtés en génération). Nous allons illustrer les ambiguïtés de la langue uniquement en analyse et uniquement sur le cas des homographes, c'est-à-dire sur deux mots qui ont la même graphie mais des sens différents (et donc généralement des traductions différentes). Ainsi, « le » est soit un article soit un pronom, « savoir » est soit un verbe soit un nom, « avocat » est toujours un nom mais soit il désigne un homme de loi (et se traduit alors par lawyer) soit il désigne un fruit (et se traduit alors par avocado). Dans la plupart des cas, les deux sens d'avocat ne crée pas d'ambiguïté réelle : dans les phrases (4) et (5) ci-dessous un humain « désambigüise » facilement ce mot par le contexte où il est employé et établit qu'avocat désigne un fruit en (4) et un homme de loi en (5).

(4) Zoé a mangé un avocat.

(5) Zoé a rendez-vous avec un avocat.

Mais pour l'ordinateur il y a une ambiguïté virtuelle introduite par le module d'analyse qui connaît les deux sens d'avocat. Cette ambiguïté virtuelle doit impérativement être levée pour éviter des traductions erronées comme (6) et (7).

(6) Zoe ate a lawyer.

(7) Zoe has a meeting with an avocado.

La levée des ambiguïtés virtuelles pour les homographes demande d'effectuer deux tâches :

— établir une classification sémantique des noms (les classes de

noms sont écrites en majuscules), par exemple : avocat est un FRUIT qui est un COMESTIBLE / avocat est un HUMAIN ;

— établir la catégorie sémantique des compléments de verbe, par exemple :

ANIME manger COMESTIBLE/HUMAIN avoir ren-~~de~~z-vous avec HUMAIN.

Mais l'affaire se complique car les verbes sont aussi souvent des homographes. Ainsi, « manger » a les sens illustrés en (4) et dans les exemples suivants :

(8) Ce poêle mange beaucoup de charbon.

APPAREIL manger COMBUSTIBLE

This stove uses a lot of coal.

(9) Les grosses entreprises mangent les petites.

ORGANISATION manger ORGANISATION

Big firms swallow up smaller ones.

(4) présente donc un phénomène d'ambiguïté croisée : pour désambiguïser « avocat », il faut désambiguïser « manger » et pour désambiguïser « manger », il faut désambiguïser « avocat ». Les ambiguïtés croisées donnent lieu à une explosion combinatoire que l'on peut chiffrer ainsi : si une phrase a n mots $m_1, m_2, \dots, m_n$, et si un mot m_i a k_i sens (donc généralement k , traductions), alors l'ordinateur doit choisir entre K hypothèses avec $K = k_1 \times k_2 \times \dots \times k_n$. On dit que le module d'analyse surgénère en produisant une prolifération d'hypothèses. Il arrive de plus qu'un homographe ne puisse pas être désambiguïsé par le contexte immédiat (la phrase où il apparaît) ni par un contexte plus large. Ainsi, en (10), on ne peut désambiguïser avocat ni par « aimer », ni par « véreux », ces deux pré-dicats s'appliquant tant à des fruits qu'à des humains.

(10) Zoé a aimé cet avocat. Pourtant, il était véreux. Zoé loved/liked this lawyer/avocado. However, he/it was shady/worm-eaten.

Soulignons le point suivant : le texte en (10) est réellement ambigu, mais il ne sera que rarement perçu comme tel en situation d'énonciation, par exemple dans un dialogue entre deux personnes car celles-ci savent bien si elles sont en train de parler des amours

de Zoé ou de ce qu'elle a mangé à midi. En TA, ou plus généralement en compréhension, on peut simuler ce type de connaissances en restreignant le domaine des textes à traiter : un système comprenant un module d'analyse ne peut produire des résultats satisfaisants que dans un domaine fermé comme le juridique, l'informatique ou l'immobilier. En effet, un module d'analyse dédié au juridique peut ne retenir que le sens « homme de loi » du mot « avocat » et ainsi limiter la prolifération d'hypothèses dues aux homographes. Cette méthode, qui n'est pas sans induire quelques erreurs, est un point de passage obligé car, rappelons-le, les homographes ne sont qu'une illustration des ambiguïtés de la langue : c'est une ambiguïté de type sémantique mais il existe d'autres ambiguïtés sémantiques (en général les temps des verbes, par exemple un présent utilisé pour un événement futur comme dans « Zoé vient demain ») et des ambiguïtés à tous les niveaux de la langue : morphologique, syntaxique et pragmatique. Cette foison d'ambiguïtés, qui débouche pour un texte sur un réseau complexe d'ambiguïtés croisées, ne peut être gérée en grandeur réelle, c'est-à-dire en simulant la compréhension humaine capable d'appréhender tant les conversations courantes que les conversations plus techniques.

Il n'en reste pas moins que les systèmes de traduction automatique progressent lentement mais sûrement avec deux tendances. Dans les laboratoires de recherche, des prototypes sont développés qui donnent des résultats assez satisfaisants dans des domaines restreints et moyennant l'utilisation d'ordinateurs puissants et un temps de calcul conséquent. Sur le marché, les produits commerciaux fleurissent. Ils produisent souvent des traductions erronées mais qui peuvent au moins servir à déterminer de quoi parle le texte (ce qui est important en veille technologique, par exemple). Ces produits couvrent un large vocabulaire, fonctionnent sur des PC et produisent instantanément la traduction. On peut penser que ces deux tendances vont converger avec le développement de la puissance des ordinateurs. Certes, mais cela n'est pas suffisant. Il reste un énorme travail à effectuer qui consiste à formaliser et à enregistrer dans l'ordinateur toutes les connaissances linguistiques et extralinguistiques, et cela ne sera pas achevé en 2001 !