SYNTHÈSE DE RAPPORT SUR L'INTERPRÉTATION AUTOMATIQUE (SPEECH-TO-SPEECH)

CONSEIL DE L'EUROPE

DGA - ITEM - SERVICE DE L'INTERPRÉTATION

CONTEXTE

Depuis plusieurs années, sous la houlette de la cheffe interprète, le Service de l'Interprétation du Conseil de l'Europe suit avec un intérêt attentif les avancées des technologies disruptives et leur impact sur l'activité de l'Organisation, d'où une réflexion précoce sur l'interprétation simultanée à distance (ISD), dans le cadre de laquelle ont été organisés, bien avant la pandémie de Covid 19, des tests d'évaluation des outils en cours d'élaboration. Grace à cette veille technologique, le Conseil de l'Europe, par comparaison avec les autres organisations internationales, avait pris un temps d'avance, ce qui a permis à son Service de l'Interprétation de proposer très rapidement, quelques semaines à peine après le début du premier confinement, une reprise des réunions avec interprétation en mode distanciel.

Ces dernières années, dans la même logique prospective, le Service de l'Interprétation a jugé pertinent, vu les progrès de l'intelligence artificielle en matière de communication multilingue (traduction automatique et projets d'extension à l'interprétation automatique, ou *speech-to-speech* – STS), de mener une évaluation des outils émergents, comme l'ont fait d'autres organisations internationales.

Cette étude, menée en déc. 2024-jan. 2025, a porté exclusivement sur la *qualité* de l'interprétation automatique en tant que telle, sans aborder pour l'instant les questions de responsabilité juridique, de sécurité informatique, de confidentialité ou de déontologie que ces nouveaux outils pourraient soulever. Il s'agit de la deuxième étude de ce type menée au Conseil de l'Europe, après un premier exercice en fév.-mars 2024.

MÉTHODE

Lors des deux séries de tests, début et fin 2024, nous avons analysé deux itérations successives d'un même outil expérimental de STS, outil que l'entreprise prestataire, avec laquelle ITEM est en contact professionnel depuis plusieurs années, a fourni au Conseil de l'Europe pour lui donner la possibilité de le tester.

Le **corpus** de discours a été choisi collégialement par les interprètes permanents du Conseil de l'Europe. Il s'agit d'enregistrements vidéo d'un éventail varié de discours prononcés lors d'événements solennels et publics organisés récemment par le CoE[1]. Les intervenantes et intervenants de rang élevé[2], locutrices et locuteurs natifs et non natifs, avec ou sans accents régionaux ou étrangers marqués, lisaient des textes rédigés ou bien s'exprimaient librement, en FR et/ou en EN.

Les **tests** ont été menés avec le soutien actif de la Direction des technologies de l'information, **en mode** purement **expérimental** : alors que les événements réels dont ont été extraits les discours avaient été interprétés par des interprètes humains, les interprétations-machine ont été produites en interne (non diffusées) et enregistrées. L'outil à tester n'a donc volontairement pas été déployé en situation réelle, du fait du risque d'atteinte à l'image de l'Organisation. Lors de la précédente batterie de tests suivis d'évaluation, début 2024, certains des discours avaient déjà fait l'objet d'une première interprétation-machine avec une version antérieure du même outil de STS.

Les **évaluateurs** étaient deux interprètes permanents du Conseil de l'Europe, l'un anglais A, l'autre français A, habitués à siéger au sein des jurys de diplôme des écoles d'interprètes. Chacun a évalué la qualité de l'interprétation vers sa langue A en utilisant la même grille d'évaluation que lors des jurys de diplôme.

Chaque discours (cf. annexes) a été **mis en page** sous forme d'un tableau à 3 colonnes : discours-source (DS) / discours-cible (DC) / observations des évaluateurs. D'autre part, chaque discours a été segmenté en tronçons de l'ordre d'une dizaine de secondes chacun (phrase ou groupe de phrases), qui ont ensuite été ventilés dans les lignes du tableau, chaque segment du DS étant juxtaposé avec le segment correspondant dans le DC et les observations des évaluateurs s'y rapportant (cf. annexes).

Le travail de **transcription des discours** a été accéléré par le recours à un autre outil d'IA (transcription automatique, ou *speech-to-text*, avec le support technique de la DIT), ce qui a toutefois nécessité de la part des évaluateurs une relecture et des corrections minutieuses avant toute évaluation, vu le nombre d'erreurs relevées dans les transcriptions automatiques. Les évaluateurs ont également noté la prosodie, tant des DS que des DC, transcrite sous forme de ponctuation ordinaire (virgules, points-virgules, points) en fonction des schémas prosodiques – montées ou descentes mélodiques – et de la durée des pauses entre deux segments parlés. Enfin, les évaluateurs ont relevé précisément dans les DS toutes les scories naturelles de l'oral (schéma prosodique inhabituel, lapsus, prononciation atypique) susceptibles de dérouter l'intelligence artificielle, là où l'auditeur éclairé et l'interprète corrigeraient spontanément.

[1] Sommet des Chefs d'État et de Gouvernement de Reykjavík (mai 2023) ; audience solennelle annuelle de la Cour Européenne des Droits de l'Homme (juin 2022) ; prononcé d'un arrêt rendu par la Cour Européenne des Droits de l'Homme dans une requête très médiatisée (avril 2022) ; séance plénière du Congrès des Pouvoirs Locaux et Régionaux (octobre 2023).

[2] Président de la République Française ; Secrétaire Générale du Conseil de l'Europe ; Président de la Cour Européenne des Droits de l'Homme ; Présidente de la Cour Européenne des Droits de l'Homme ; Président du Congrès des Pouvoirs Locaux et Régionaux.

Les **critères d'évaluation** ont été les mêmes que pour l'évaluation de candidats humains lors d'examens de diplôme dans les écoles d'interprètes, avec une comparaison fine (cf. explications détaillées *infra*) des discours-cibles (interprétation automatique) avec les discours-sources (discours originaux enregistrés, produits par des locutrices et locuteurs humains lors d'événements réels), avec comme point de référence l'interprétation que l'on attend de la part d'une ou d'un interprète humain professionnel.

En cas de doute sur la pertinence de tel ou tel jugement ponctuel, les évaluateurs se sont systématiquement reportés au prononcé (enregistrement), dans la mesure où c'est la **prestation orale** que l'on évaluait et non pas sa transcription écrite, cette dernière servant uniquement de soutien visuel au travail d'analyse détaillé.

ANALYSE GLOBALE

L'**impression superficielle** est, de prime abord, **plutôt positive**.

Par rapport à l'itération précédente, le **message semble mieux rendu**, tant dans son contenu que dans sa forme. Les **erreurs manifestes** de traduction sont **moins flagrantes**, on note même un certain nombre de **restitutions précises et élégantes**, dans les deux langues, à l'échelle d'une phrase voire d'un paragraphe entier.

Enfin, lorsque le débit est normal, la **voix** est beaucoup plus **agréable** : elle n'a plus le caractère robotique de l'itération précédente et l'on entend maintenant une **prosodie**[3] souvent **naturelle**, analogue à celle d'un être humain.

Toutefois, on est immédiatement frappé par la **longueur excessive des pauses** (fréquemment de l'ordre de 6 secondes mais pouvant aller jusqu'à la vingtaine de secondes – les pauses naturelles des locuteurs humains sont en général inférieures à la seconde) entre les phrases voire en milieu de phrase, ce qui est **déroutant** et **désagréable**.

Vraisemblablement pour rattraper le retard accumulé, le débit **accélère** alors parfois **brutalement**, là où une ou un interprète s'efforcerait de lisser son débit.

Outre ces **défauts prosodiques immédiatement apparents**, l'impression initiale positive ne résiste pas à un examen attentif : une analyse détaillée fait apparaître de **nombreux défauts structurels récurrents**. Ils sont repris ci-dessous point par point, quitte à revenir brièvement sur certains évoqués ci-dessus.

[3] *Prosodie* : ensemble des traits rythmiques et mélodiques du discours : pauses, débit, ralentissements ou accélérations, montées ou descentes mélodiques

ANALYSE DÉTAILLÉE

1. PROSODIE ARTIFICIELLE

On note des **pauses trop longues**, suivies d'un **passage brutal** à un **débit extrêmement rapide**.

L'éventail des vitesses paraît **discontinu** : la machine semble ne disposer que d'une palette restreinte de vitesses possibles (1ère / 2ème / 3ème) entre lesquelles elle embraie, donnant parfois des embardées désagréables pour passer sans ménagement de la première au turbo, ce qui est non seulement **déstabilisant** ou **risible** mais rend le message **incompréhensible**.

De ce fait, la prosodie semble par moments **artificielle** ou **mécanique** par rapport à un orateur humain, qui modulerait son débit de manière fluide.

2. DÉCOUPAGE SÉMANTICO-SYNTAXIQUE ABERRANT

Les pauses prosodiques sont non seulement trop longues mais souvent étrangères au discours original. Dans certaines phrases complexes, le **découpage prosodique,** entendu par l'auditeur comme un **découpage sémantico-syntaxique**, est fréquemment **erroné**.

En français et en anglais, les compléments circonstanciels (compléments nominaux ou propositions subordonnées circonstancielles) peuvent être placés en tête ou en fin de phrase. C'est à l'auditeur de repérer correctement, sur la base de la prosodie et du sens global du message, si tel ou tel complément circonstanciel est à rattacher à la proposition qui précède ou à celle qui suit. Il n'y a, en général, aucune ambiguïté pour l'interprète, grâce à l'analyse du sens à laquelle elle ou il se livre.

Dans les restitutions par le STS, les compléments circonstanciels qui étaient clairement en fin de phrase dans le DS sont presque systématiquement rattachés par erreur à la phrase suivante dans le DC : **faux sens** voire **non-sens**.

Inversement, il peut se produire que des compléments circonstanciels qui étaient en tête de phrase dans le DS soient rattachés par erreur, dans le DC, à la fin de la phrase précédente : erreur symétrique de la précédente, aux conséquences analogues.

Certes, le STS est parfois capable de "saucissonnage" efficace, lorsqu'une phrase complexe du DS est débitée à bon escient en tronçons autonomes dans le DC tout en en respectant rigoureusement le sens du message, comme le ferait une ou un interprète. Toutefois, à l'échelle d'un discours entier, on note dans le DC davantage de découpages erronés que justifiés, ce qui **déforme** fréquemment **le propos** voire rend le **message incompréhensible**.

3. REPÉRAGE ERRONÉ DES ANAPHORES[4]

L'**anaphore** est fréquemment **mal repérée**, et donc **mal ou pas restituée**.

En français, on désambiguïse fréquemment l'anaphore par des marques grammaticales telles que le *nombre* (singulier / pluriel) ou le *genre* (masculin / féminin), qui peuvent être audibles à l'oral (articles, pronoms, terminaisons, liaisons phonétiques, genrés ou nombrés). Toute erreur d'anaphore audible est susceptible de **brouiller le sens**.

Exemple d'anaphore externe mal repérée et audible : erreur sur le genre d'une personne présente en salle qui, dans le DS, avait été citée uniquement par son titre épicène (non genré) dans l'original anglais (*the Commissioner*). Ainsi, *la* précédente Commissaire aux Droits de l'Homme – *the Commissioner for Human Rights* dans le DS - est devenue dans le DC *le* Commissaire aux Droits de l'Homme. Cela aurait été **insultant** pour la personne concernée et aurait **risqué de porter atteinte à l'image de l'Organisation**.

Autre erreur d'anaphore : un même terme dans le DS (*the Court*) est rendu à deux phrases d'intervalle dans le DC par *la Cour* (CEDH) puis par *le tribunal* (juridiction interne à un État donné – sachant que le mot *court avec une minuscule* peut souvent être rendu par *juridiction*), alors que, dans le DS, d'une part *the Court* était à entendre chaque fois sans aucune ambiguïté comme signifiant *la Cour Européenne des Droits de l'Homme*, d'autre part, dans son argumentation, l'orateur opposait clairement la Cour Européenne des Droits de l'Homme aux juridictions internes. Ce **contresens** aurait **risqué de porter atteinte à l'image de l'Organisation**.

Erreur analogue : à une autre occasion, *le Conseil* (sous-entendu : *de l'Europe*) dans le DS est rendu dans le CD par *the advice*, les divers référents possibles du mot *Conseil* n'étant pas correctement discriminés. Cette **erreur risible** aurait **risqué de porter atteinte à l'image de l'Organisation**.

À rattacher à ce phénomène : la capacité du STS à se contredire, à deux phrases d'écart (cf. manque d'analyse identifié *infra* au point n° 8). Le DC produit par le STS peut être entaché d'**incohérences grossières**.

4. LIAISONS PHONÉTIQUES FAUTIVES

La réalisation des **liaisons phonétiques** est aléatoire : parfois à bon escient, mais souvent elles sont soit manquantes là où on les aurait attendues, soit superflues, ce qui choque l'oreille et peut même faire douter l'auditeur du sens voulu, par exemple là où on entend un singulier alors qu'on attendait un pluriel que la liaison obligatoire aurait rendu audible (cf. anaphores erronées évoquées *supra*).

5. INCAPACITÉ GLOBALE À GÉRER LE MÉTA-DISCOURS ET LE MÉTA-MESSAGE

[4] *Anaphore* : renvoi
– soit interne, vers un autre élément du discours, un repérage correct de ce dernier au sein de ce même discours ;
– soit externe, vers l'univers extra-discursif, bien que celui-ci n'ait pas été cité expressément dans le discours.
Exemples : utilisation d'un pronom personnel pour renvoyer à un groupe nominal déjà cité, ou utilisation successive de deux synonymes pour renvoyer à une même notion (phénomène fréquent en français pour des raisons stylistiques).

À un niveau plus abstrait (rhétorique et discursif et non plus syntaxique), le STS, du fait de son manque inhérent de capacité d'analyse, semble incapable de gérer le **méta-discours**[5]. Cela se manifeste de diverses manières :

- **Incapacité à distinguer le méta-discours du discours lui-même.**

  Notamment, le STS ne parvient pas à gérer les énumérations, par exemple lorsque le locuteur numérote explicitement (en mode méta-discursif) les éléments successifs d'une liste, par exemple les diverses dispositions d'un arrêt rendu par la CEDH *: 1, 2, 3, ...* Ces chiffres, qui font clairement partie du cadre *formel* de l'énoncé et non pas de l'énoncé *lui-même*, sont perçus et restitués par le STS comme faisant partie intégrante des phrases qu'ils délimitent, ce qui donne des énoncés erronés voire risibles. Ainsi, dans le DS, « *6. Judges that* etc » (du verbe *to judge*, dont le sujet sous-entendu, *The Court*, avait été énoncé en tout début de liste pour être mis en facteur commun d'une série de verbes dans des alinéas successifs), et ce, malgré la pause prosodique marquée entre « *six* » et « *judges* » dans le DS et malgré le schéma prosodique particulier propre à ce type d'énoncé, devient dans le DC le syntagme nominal « *six juges* », compris par l'auditeur comme « *six magistrats* » ! Le STS ne "comprend" pas que ces chiffres s'appliquent aux parties du discours à un niveau formel, c'est-à-dire à leur rapport structurel les unes aux autres, mais cherche à rendre ces chiffres comme faisant partie du plan du discours, comme si c'étaient des mots faisant partie de la phrase énoncée elle-même, d'où un **non-sens**.

  Inversement, des signes graphiques, qu'il fallait comprendre dans le *sens* qu'ils prennent dans le contexte de la phrase, sont parfois énoncés dans leur *matérialité graphique*, là où l'être humain saurait instantanément à quel niveau de lecture il faut se situer.
  Ainsi, dans le prononcé de l'arrêt *Klimaseniorinnen*, là où le DS parlait de *preindustrial level* et où, dans le DC, on attendait simplement l'équivalent ordinaire *niveau pré-industriel*, le DC devient : *niveau P-R-E* (trois lettres épelées séparément) - *accent aigu* - *industriel* [sic]. On peut supposer que c'est la synthèse vocale (*text-to-speech*) français-français, dernière étape du processus en trois temps[6] sous-jacent au STS, qui pèche, ne "sachant" pas s'il fallait lire l'assemblage de lettres *pré* en tant que mot ou bien l'épeler comme si c'était un sigle – arbitrage qui met en jeu une capacité intuitive de discernement qui relève de l'aptitude, proprement humaine, à distinguer entre le discours et le méta-discours. **Non-sens**.
  De même, le chiffre *trente mille,* dans le DS du maire de L'Haÿe-les-Roses, devient dans le DC « *thirty zero zero* » (**non-sens**), la synthèse vocale n'ayant vraisemblablement pas su interpréter au sens fort du terme le signe graphique *3000* (notons au passage qu'il manque un zéro dans le DC par rapport au DS : il y aurait eu de toute manière un **faux sens**) qui avait été produit par la traduction-machine, étape intermédiaire du processus de STS.

- **Incapacité à distinguer les expressions figurées**, d'où des traductions parfois trop littérales.

[5] On parle de *méta-discours* lorsque le discours se prend lui-même pour objet, phénomène extrêmement fréquent (« Si je dis cela, c'est pour etc… ») et qui peut prendre des formes diverses.

[6] 1° reconnaissance vocale en langue source (speech-to-text, STT)
2° traduction automatique de langue-source en langue-cible (traduction-machine, ™) de texte à texte
3° synthèse vocale en langue cible (text-to-speech, TTS)

S'il peut arriver que certaines expressions figurées consacrées, dans le DS, soient rendues dans le DC avec pertinence voire élégance par d'autres expressions figurées consacrées (expressions figées de la langue-cible qui sont des équivalents statistiquement attestés de celles de la langue-source dans un contexte lexical donné, car figurant dans le corpus considérable sur lequel se fonde le STS), en revanche les expressions imagées du DS qui ne sont *pas* de simples expressions figées mais qui sont le fruit de l'invention personnelles du locuteur et qui, pour être inhabituelles, n'en sont pas moins riches de sens, sont rendues platement à la lettre dans le DC si bien que **le message ne passe pas**. Or le discours figuré, ou imagé, relève également du méta-discours car il se présente implicitement comme étant à ne pas prendre au pied de la lettre mais demande à être interprété, au sens fort du terme, ce que sait faire l'être humain mais non la machine.

6. MAUVAIS DÉCOUPAGE DES GROUPES NOMINAUX COMPLEXES

Quand on interprète par exemple de l'anglais vers le français, les **syntagmes nominaux complexes**, fréquents dans le discours technique ou juridique (de type N1 N2 N3 … Nn, où chaque nom de la série doit être compris comme qualifiant le nom ou le groupe nominal qui le suit), nécessitent que dans le DC on inverse l'ordre de tous les mots pris un par un. Cela suppose la mise en mémoire d'informations multiples pour une restitution dans l'ordre inverse de celui où elles ont été entendues, ce qui est parfois un défi pour l'interprète lorsque la série est longue.

Ces syntagmes nominaux complexes sont souvent **segmentés abusivement** par le STS. Au lieu de d'inverser l'ordre de tous les éléments d'information et surtout de les hiérarchiser correctement, le STS scinde souvent de manière arbitraire les derniers éléments du syntagme complexe et les traduit par un deuxième groupe nominal autonome, ce qui non seulement produit une **phrase agrammaticale** mais provoque un **non-sens**.

Tout comme les métaphores inventives évoquées précédemment, les syntagmes nominaux complexes peuvent relever de l'invention individuelle dans une situation donnée (le phénomène est particulièrement marqué en allemand où tout locuteur est libre de créer des néologismes sous forme de noms composés). C'est cette créativité individuelle, propre à un locuteur donné dans des circonstances données, à laquelle le STS est incapable de se mesurer.

En effet, si le STS fonctionne sur la base de corpus extrêmement volumineux (millions ou milliards d'items, bien plus que ce que peut connaître et mémoriser un être humain), il dépend néanmoins d'occurrences déjà attestées. C'est notamment la raison pour laquelle, comme cela a été évoqué *supra*, les formules figées du DS sont parfois restituées avec beaucoup d'élégance dans le DC par d'autres formules figées. Si cela peut impressionner de prime abord, il convient de rappeler que cette partie « émergée », « flatteuse », du processus d'interprétation est néanmoins mécaniques, elle ne demande pas de réflexion au sens fort du terme, y compris chez l'interprète, car elle relève du simple bagage lexical et d'un peu de savoir-faire, lequel peut dans certaines circonstances fonctionner en pilote automatique chez l'être humain, donc être automatisé et confié à une machine. En revanche, les métaphores ou les associations de mots inventives, singulières, sont mal rendues par le STS : elles sont traduites littéralement car le STS ne peut se rattacher à aucune occurrence préétablie attestée dans le corpus de langue source. Le résultat de cette traduction littérale est que, souvent, **le sens ne passe pas**, là où l'interprète saurait le restituer efficacement.

En somme, le STS est par sa nature même incapable de gérer les événements singuliers. Il est capable de traiter, parfois avec brio, un discours formaté, prévisible, qui, pour aussi brillant qu'il soit, accumule les clichés (*langue de bois*, inévitable dans certaines circonstances de la vie sociale), mais dès qu'il y a individualité et singularité langagière, il accuse ses limites.

7. INCAPACITÉ À CORRIGER LES SCORIES DU DISCOURS

En général, le STS **a du mal à rectifier les erreurs ou maladresses dans le DS** (lapsus, déformations phonétiques, ruptures de construction, tournures malhabiles – qui sont une autre forme de la singularité du locuteur) afin d'en extraire un sens cohérent et de le restituer – ce qu'en revanche sait faire intuitivement tout interprète professionnel.

8. RÉPÉTITIONS ABUSIVES DE SEGMENTS ENTIERS

Dans certains passages à structure syntaxique et/ou rhétorique complexe, une idée qui, dans le DS, avait été énoncée une seule fois est, dans le DC, énoncée une première fois en début de phrase et répétée une deuxième fois sans raison valable en fin de phrase ou de paragraphe. Cette **mauvaise gestion des éléments d'information** est un cas particulier des ***hallucinations*** ou ***confabulations*** de l'IA.

Lorsque l'idée est répétée à l'identique, cela donne au mieux une **répétition** (Spano 10 et 16), au pire une **lapalissade**. Si, en revanche, l'idée est répétée en ajoutant la deuxième fois une négation, le discours devient alors **contradictoire** voire **incohérent** (délirant).

Quel que soit le cas de figure, le DC **perd toute crédibilité** auprès de l'auditeur.

* * *

À la lumière des divers point faibles analysés précédemment, le STS manifeste une **absence de vérification de la cohérence interne** du DC (capacité à se répéter en mode écholalique, ou à dire une chose et son contraire, parfois à deux phrases d'écart), qui serait rédhibitoire chez une ou un interprète dont la qualité première est l'auto-vérification constante. Cette **absence de contrôle-qualité** est imputable à l'absence d'analyse par le STS : contrairement à l'être humain, **la machine ne pense pas**, elle est notamment incapable de se remettre en cause et de se corriger, tandis que l'interprète porte constamment un regard critique sur sa propre prestation et sur ses propres processus cognitifs. De même que la machine est incapable de mentir, elle est, à la différence de l'être humain, incapable de dire lorsque c'est nécessaire : "L'interprète se corrige".

SYNTHÈSE

Le STS, en l'état actuel, peut faire superficiellement illusion car, avec une prosodie qui se rapproche désormais de la prosodie humaine naturelle, il est capable de produire ponctuellement des phrases complexes, souvent

bien structurées d'un point de vue grammatical et rhétorique, qui emploient un vocabulaire riche, précis et technique.

Toutefois, la durée excessive des pauses (silences), en fin voire en milieu de phrase, amène souvent la machine à accélérer brutalement, d'où un débit par moments difficilement compréhensible ou supportable, là où les interprètes humains sauraient, sans modifier exagérément leur débit de parole, synthétiser le propos à bon escient lorsque l'oratrice ou l'orateur est prolixe et très rapide ou lorsqu'ils ou elles ont elles-mêmes pris du retard.

En outre, un examen attentif des productions du STS révèle qu'on a affaire à une espèce de machine folle : si elle est certes capable par moments de productions cohérentes, émaillées de trouvailles étonnantes, elle commet aussi nombre d'erreurs grossières que ne commettrait pas un être humain. Si certaines erreurs sont immédiatement apparentes, les plus pernicieuses sont les erreurs discrètes, susceptibles de passer inaperçues, du fait de l'apparence de prime abord agréable du discours-cible, mais faussant radicalement le sens du message (ce que les interprètes appellent un *discours parallèle*, rédhibitoire en situation professionnelle ou d'examen).

A l'heure actuelle, les **principaux points faibles** sont :

- **la longueur excessive des pauses prosodiques (silences), les transitions brutales, « mécaniques », entre les divers débits de parole, parfois trop rapides pour être compréhensibles ;**

- **l'incapacité à découper correctement les phrases du discours-source (compléments circonstanciels rattachés à la mauvaise phrase ou au mauvais verbe : *inability to parse)*, par incapacité à analyser le sens profond du discours, d'où des faux sens quasi systématiques ;**

- **le manque de cohésion et de cohérence, par incapacité à gérer les anaphores (erreurs de renvoi interne au discours, d'où des contradictions internes parfois criantes ; ou erreurs de renvoi vers le monde extra-linguistique, d'où des propos absurdes ou en contradiction flagrante avec la réalité environnante).**

- **l'incapacité à gérer le métadiscours et le discours figuré : l'IA, contrairement à ce que son nom pourrait laisser penser, ne « réfléchit » pas ; elle se contente de traiter le discours en s'appuyant sur un corpus statistique certes considérable mais qui reste très en-deçà de l'infinie capacité d'invention et d'interprétation des locuteurs humains.**

- **de plus, le corpus du STS, tiré d'occurrences écrites, ignore le champ considérable des expressions de la langue parlée, ce qui peut le mettre en échec. Certaines tournures parlées, ou relevant de l'invention propre au locuteur, sont rendues littéralement, si bien que le sens ne passe pas, ou pas immédiatement, ce qui est tout aussi gênant car le message oral, du fait de sa nature évanescente, doit être immédiatement compréhensible, sous peine de ne pas être compris du tout.**

Bien qu'il soit difficile pour des interprètes, non-spécialistes de l'intelligence artificielle, de déterminer avec exactitude les facteurs auxquels sont imputables ces points faibles, d'autant que l'outil fonctionne parfois, y

compris désormais pour les spécialistes, comme une *black box* dont on se contente de constater les productions dans une situation donnée, il est néanmoins permis d'avancer les **hypothèses suivantes**.

Là où l'interprète procède par une analyse du sens du message qui fait appel à la capacité de réflexion proprement humaine (mise en parallèle et mise en opposition implicite des éléments d'information ; va-et-vient constant entre ce qui est dit, ce qui a déjà été dit et ce qui pourrait être dit par la suite ; saisie des liens de cause à effet implicites entre les éléments ; liens implicites entre les éléments énoncés et le contexte extra-discursif ; mais aussi retour réflexif constant sur ses propres processus cognitifs, qui joue le rôle d'un contrôle-qualité draconien), le STS quant à lui procède à une analyse prédictive à très grande échelle, sur la base de corpus statistiques considérables, et non pas sur la base d'une analyse et d'un raisonnement, contrairement à l'être humain.

Ce n'est pas la méthode que nous évaluons ici mais les prestations : la question est de savoir si le STS, avec les méthodes qui lui sont propres, est capable de parvenir à un résultat de qualité égale à celui que produit une ou un interprète.

Ce que nous savons par ailleurs, c'est qu'à l'heure actuelle le STS procède en 3 étapes successives, quoique quasiment instantanées :

- reconnaissance vocale (transcription automatique du prononcé), ou *speech-to-text* (STT) en langue-source
- traduction automatique de cette transcription, ou *machine translation* (MT), avec passage de la langue-source en langue-cible
- synthèse vocale de la traduction, ou *text-to-speech*, en langue-cible

Si l'étape finale de synthèse vocale (*text-to-speech* ou TTS) ne semble pas en cause en tant que telle (par exemple, il est peu probable que les pauses erronées ou trop longues soient produites de manière autonome par la synthèse vocale, d'autant que, curieusement, hormis les pauses fréquemment trop longues et mal venues, la prosodie du DC est par ailleurs globalement correcte voire naturelle) ; et s'il est de même peu probable que ce soit l'étape intermédiaire de la traduction-machine qui se trompe dans le découpage syntaxique, et qui rattache par erreur des compléments circonstanciels à la mauvaise phrase en langue-cible, alors que les signes de ponctuation, et donc la structuration des éléments d'information, ont été fournis par la transcription automatique du DS ; c'est dès lors vraisemblablement l'étape initiale de la transcription-machine du DS (*speech-to*-text en langue-source) qui pèche : si elle produit une analyse syntactico-sémantique erronée, cela se traduit dans la transcription du DS par des découpages et regroupements aberrants d'éléments de sens, et donc, par effet de cascade, cela génère des erreurs dans la traduction-machine, et, en bout de chaîne, dans la synthèse vocale en langue-cible, autrement dit dans le DC.

Il serait intéressant de voir ce que produirait un système de STS auquel on aurait **fourni le texte du discours original**, tout comme on fournit idéalement aux interprètes le texte d'une intervention destinée à être lue – mais cela reviendrait à faire produire une traduction automatique, suivie éventuellement d'une synthèse vocale, ce qui ne présente pas grand intérêt pour les situations réelles de communication où l'on a recours à l'interprétation. Dans ces conditions, on peut se demander si l'étape de la synthèse vocale est vraiment nécessaire, hormis pour un public malvoyant, et s'il ne serait pas plus simple de proposer simplement un simple

sous-titrage du discours dans telle ou telle langue-cible (*speech-to-text* avec changement de langue), avec toutes les limites reconnues que comporte la traduction-machine non corrigée par un être humain, sans compter, pour les lectrices et lecteurs, la fatigue visuelle et cognitive qu'engendre un texte qui s'auto-corrige constamment à l'instant même où il s'affiche, comme c'est le cas pour la plupart des outils de ce type.

Toutefois, cela ne résout pas le défaut majeur du STS qui est l'**absence d'analyse / absence de prise en compte du méta-discours**. C'est manifestement un point faible structurel de l'outil, alors que cette prise en compte du méta-discours et du méta-message, et la capacité d'analyse fine qui la sous-tend, sont des éléments essentiels du travail de l'interprète. En l'état actuel, un jury de fin d'études, composé de professionnels, ne décernerait pas le diplôme d'interprète à l'outil de STS.

Il semble donc que, malgré des améliorations superficielles notables, quoique imparfaites, qui ont essentiellement trait à la mise en forme sonore du message, le STS présente encore des **points faibles structurels** tenant à son mode de fonctionnement, qui **faussent ou entravent la communication de manière rédhibitoire pour les usages auxquels est destinée l'interprétation de conférence, et risquent par ailleurs de porter atteinte à l'image de l'Organisation.**

A défaut d'envisager le recours à l'interprétation de conférence automatique dans un avenir proche au Conseil de l'Europe, il convient en revanche d'envisager la ou les manières dont les interprètes pourraient s'appuyer sur l'intelligence artificielle dans l'exercice de leur tâche, en utilisant ces nouveaux outils de manière éthique et responsable pour rendre l'interprète plus efficace encore. C'est là tout un champ à explorer, qui intéresse vivement les interprètes. Par exemple, l'intelligence artificielle, utilisée à bon escient, pourrait peut-être permettre aux interprètes de mieux se préparer encore pour leurs réunions, en traitant rapidement et efficacement des volumes élevés d'information écrite. Toutefois, cela soulève des problèmes importants de confidentialité et de déontologie, d'autant que les réunions mettant en jeu des documents écrits complexes et volumineux sont souvent de nature sensible et hautement confidentielle (CPT, GRECO, Comité gouvernemental de la Charte Sociale, Cour Européenne des Droits de l'Homme…).

Quoi qu'il en soit, c'est là un domaine d'études vaste et prometteur, pour lequel il sera, d'une part, souhaitable de recueillir les propositions et l'avis des interprètes eux-mêmes, et d'autre part, vraisemblablement utile de faire appel, à titre ponctuel voire pérenne, à des spécialistes à l'interface des deux disciplines (interprétation ET intelligence artificielle), tels que commencent à en former un certain nombre d'établissements d'enseignement supérieur et de recherche en France et à l'étranger.