Report on WORDLY AI Interpretation

Introduction

This report was initiated and conducted by the WHO interpretation team (INT) as a response to requests from WHO Technical Units (TUs) for recommendations and advice on the use of AI interpretation as a means to maintain multilingualism at meetings where funds were short.

After an initial assessment a posteriori of a few cases where AI interpretation was used and, particularly noting that it was not of sufficient quality to allow for use in WHO meetings and that it involved a significant reputational risk for the organization and for speakers at such events, INT decided to conduct a thorough study on AI interpretation in all 6 official languages. For that purpose, among others, an interpretation intern was recruited to assist in the process.

While AI interpretation involves numerous elements that require in-depth examination such as cost and technical integration with meeting systems as well as legal accountability, IT security, confidentiality and ethical issues related to inherent bias and the quasi-monopoly on AI sources, it was decided to limit the study to the field of expertise of INT and solely assess the quality of AI interpretation and reputational risks involved. The study also aims to set a baseline for future assessments of AI interpretation in the 6 languages, as it develops over time.

Description of the process from selection to assessment

The first task was the **selection of the AI interpretation provider** to be tested. Our AI interpretation research started with a market study of available AI tools. It quickly became clear that many platforms only provided AI-assisted interpretation, that is AI tools to assist human interpreters in their work, and not AI interpretation proper. Others provided only part of the cascade triad involved in AI-interpretation, i.e. speech-to-text (STT), text-to-text through Machine Translation (MT) and text-to-speech (TTS). Those platforms were excluded from the scope of the study as they did not meet the needs expressed by TUs. It was also decided to exclude platforms that did not cover the 6 official languages of WHO. Finally, Interprefy was excluded, as it had recently been assessed by WTO. The 2 remaining AI interpretation providers that met all conditions were KUDO and WORDLY. KUDO did not respond to our request for a test. WORDLY responded quickly and diligently. Further, it was considered as the top provider of AI interpretation by NIMDZI, the external consultants tasked with the functional review of WHO language services (LNG). It was therefore decided to go ahead with WORDLY. Since the selection of WORDLY, KUDO was tested by another International Organization, and the results of the test are very similar to the conclusions reached by WHO INT on WORDLY.

INT developed specific **speech selection criteria**. They include, among others, accents, numbers, acronyms, figures of speech, cultural references and speed. We chose for each language 3 speeches involving various difficulties to be tested. We decided to use speeches from the 2024 World Health Assembly as the recordings were public and readily available and there was a sufficient variety in all 6 languages. A more detailed explanation can be found in the annex entitled Speech Selection Process.

As we quickly realized that a major issue with WORDLY was the identification of the language spoken, we decided to test in one multilingual series 6 speeches in 6 languages. In addition, 2 speeches per language were tested into all 5 other languages separately, where the source language was indicated

from the very beginning and WORDLY did not need to identify it, so that the initial difficulty with identifying the source language did not impact the assessment in all cases.

Quality assessment criteria were developed and fine-tuned, based on the UN competitive exam assessment criteria in a simple division into content, expression, and delivery, and a 3-tier grading of good (1), poor (0.5) and unacceptable (0). We adapted the criteria to the specificity of AI interpretation. We, thus, decided to focus in the grading on content, taking expression and delivery into account only insofar as the meaning was impacted. We decided to grade each speech by segment, as divided by WORDLY in the AI transcription, to facilitate and standardize the grading of all speeches. 75% was set as the passing grade. The assessment criteria can be found in the annex entitled Quality Assessment Criteria for AI Interpretation.

The **assessment** was conducted by the LNG staff, in both INT and Translation (TRA), and with the assistance of one current and one former interpretation teachers from the University of Geneva, Faculty of Translation and Interpretation (FTI).

A file was created by the INT intern for each speech, including the audio of the original speech, the audio of the output and a template for assessment. Each template had a transcription of the original and the output to facilitate grading. The transcript of the output was based on the WORDLY AI transcription but had to be adjusted to the audio output as there were significant differences. This was not done in all cases, but the grading was always based on the audio output.

Speeches were distributed amongst the graders for assessment according to language combination and availability, after explanation of the process and preparation of the templates. Each template was divided into segments of 1 or 2 sentences, based on the division in the AI transcription. Each segment was graded 0, 0.5 or 1. A total grade was added at the end of each template and translated into a percentage, with a brief comment on the interpretation of each speech.

For speeches where none of the graders had knowledge of both the Source and Target Languages, an English translation was used as relay.

There was one blind assessment, that is an assessment of the output without checking the source. The Chinese speech in the multilingual series was assessed in all languages by non-Chinese speakers, to check the pure comprehensibility of the output.

Each grader was asked to identify **reputational risks** for each assessed interpretation. These related to elements in the interpretation that could threaten the values, image or identity of WHO or of the speaker and their country or organization; that could ridicule the speaker; that could have political or diplomatic fallout; or that could undermine the smooth functioning of the meeting. A single reputational risk was considered as eliminatory.

Assessment results

The assessment results are surprisingly low for all languages. They range from 5% to 83% with only 1 interpretation out of 90 getting a passing grade. Not a single interpretation was free of reputational risks which ranged from 1 to 9 in a single speech. The overall average is 46%, with interpretation into English at 51% faring better than other languages but, interestingly, not by a huge margin. Even more interestingly, interpretation into English had the highest total number of reputational risks at 46 RR. Interpretation into Chinese had, at 40%, the lowest average grade. Interpretation from English was also the highest graded at 54% but, surprisingly, French as a source language had the lowest average

at 36%. This may be due to the very high difficulties in the French texts, particularly the one in the multilingual series which got the overall lowest grades.

While, as expected, the multilingual series fared, at an average of 40%, less well than the bilingual one, which averaged 48%, it was, again, not by a wide margin. The choice of speeches clearly had a significant impact. Nevertheless, the test clearly showed difficulties in language identification and **code-switching**. Other than taking the time of a sentence or two to switch from one language to another, WORDLY shadowed the English-speaking Chair, **interpreting** him **from English into English**, and often inaccurately.

In terms of dealing with specific **difficulties**, AI interpretation did well with **speed**. However, high speed did affect completeness. There was a significant **time lag** between the beginning of each speech and the beginning of interpretation as the AI cascaded through the triad process, including contextualization and auto-correction, and produced the required output. This lag extended at times to over 32 seconds while in human interpretation, it is usually no more than 5 seconds. While that is perturbing for the listener, the bigger problem is that, as a result, the last sentences were not always interpreted.

One difficulty that was more challenging for AI was **proper nouns**, including names of countries and names of people. Training could possibly help but we were told by WORDLY that it was not possible to train their AI interpretation tool. While a human interpreter will use context to navigate the difficulty or will circumvent it by omitting for instance the unfamiliar name of a President and simply using their title, AI interpretation does neither. The most distinct examples are “Brunei Dar Essalam”, that was interpreted from Chinese into Arabic as “the brunette Russel”, Greece as Chris and Haiti as Heidy in all languages. Also, Dr Moeti, the AFRO Regional Director was misgendered as a man (inherent gender bias) and interpreted as “our African” in Arabic. This seems to be an issue with **ST.** Even more seriously, when Hamas was referred to as having perpetrated terrorist attacks in the statement of Spain, it came out as an incomprehensible “Ifer” in the Arabic interpretation, (Ifer does not mean anything) while the AI transcription mentioned the US instead of Hamas. Such errors are serious reputational risks as they ridicule the speakers and the countries involved and could even cause serious diplomatic incidents.

This was also the issue when a foreign language expression was used as a **cultural reference**, for example, in the statement of Bangladesh. The speaker exclaimed at the end of his speech “Joy Bangla”, the national slogan of Bangladesh in Bengali, that translates as Hail Bengal. While a human interpreter, unfamiliar with the words would have either repeated them as is, or omitted them if unsure, AI interpretation used Joy Bangla as the name of the Chair and misgendered him as female, in languages that gender nouns, based on the assumption that this was a woman’s name. This seems to be an issue with **MT**. The gross error is of course unacceptable, as it ridicules the speaker and could be deemed offensive to the Chair.

Figures were also a major stumbling-block for AI. Some came out correctly. Many were incorrectly transcribed and pronounced, especially when there were many zeroes involved and even in dates. This leads us to think the issues are in **STT and TTS**. The resulting output made the relevant sentences incomprehensible, and made the speakers sound incoherent.

Complex grammar and syntax were more problematic in some languages than others, particularly from Arabic. A notable example was in one speech, where the speaker mentioned the reduction in maternal mortality “by about 70%” which was translated as “to about 70%” in French and Russian.

Technical terms were also problematic**.** Transmission of polio was interpreted from Arabic as transportation, due to the similarity of the 2 words in Arabic. In a statement in French, hepatitis became Ebola in Arabic. From Chinese "stratified health" was interpreted into all languages as "airplane health".

While the grades, as mentioned above, were only for content, it is important to note that **expression** was often poor and literal.

Delivery, while also not graded, was overall better than other platforms and did not suffer from sudden rapid acceleration. However, it was extremely monotonous and unexpressive making it difficult to follow for more than a few minutes. **Pronunciation** in Arabic and Russian was incorrect at times. In Arabic, wrong vowels changed the meaning of certain words; in Russian, the wrong syllable was emphasized in some words, making comprehension difficult.

Conclusions and Recommendations

The grades for AI interpretation for all language combinations ranged from 5% to 83% and the number of reputational risks per speech ranged from 1 to 9. Only one interpretation out of 90 got a passing grade. It was English into French. However, all interpretations of all speeches had at least 1 reputational risk. The **conclusion** is that AI interpretation is still at an experimental stage and is not fit for use in WHO meetings with external stakeholders, as is currently stipulated in the recent guidance on the use of AI in the workplace, issued in Information Note 2025/3. In line with that guidance, AI interpretation may be used in internal meetings involving WHO staff only. Where AI interpretation is used under the aforementioned conditions, it is **recommended** that staff who understand the languages used be present to avoid major miscommunication. It is also **recommended** that recordings be made and sent to INT for assessment and monitoring purposes.

In view of the progress noted in AI interpretation in the last 2 years and, potentially, in upcoming years, it is **recommended** that resources be allocated to the continued monitoring and assessment of AI interpretation, particularly through the recruitment of interpretation interns and cooperation with academic institutions, such as the FTI. It is also recommended to explore other avenues for cooperation on AI language services, other than AI interpretation.

It is worth noting that this is a **baseline study** to be used in future comparative assessments by WHO and possibly other organizations and institutions.

This study may, in the future, be built on by FTI which would use its research resources to draw more specific conclusions on AI interpretation in the various languages and issue an **academic paper** on its basis later on.

While this study is limited to the assessment of the quality of AI interpretation, issues with **technical system interoperability, IT security**, **confidentiality**, **dependence on quasi-monopoly of AI**, political issues regarding the **source of AI**, **legal accountability**, **inherent bias** (as to gender and race among others), **hidden cost** and **carbon footprint** are some of the numerous issues to be further studied before making an informed decision on the use of AI interpretation.

ANNEX 1 – Speech Selection Process

The input that we decided to use are speeches from the 77th World Health Assembly (WHA) held in May 2024. This is public material that does not undermine confidentiality and involves a mix of political statements and technical health-related content.

We took into account geographical representation:

SPEECHES			Percentage of population
EUROPE 16.6%	Eastern Europe	2	9.11%
	Western Europe	1	
ASIA 44%	East Asia	3	57.8%
	Central Asia	1	
	West Asia (Middle East)	3	
	South Asia	1	
AFRICA 16.6%	East Africa	1	18.5%
	Central Africa	1	
	West Africa	1	
AMERICAS 16.6%	North America (Caribbean)	1	13.1%
	South America	2	
OCEANIA 5%	Australasia	1	0.6%

We identified elements that constitute a difficulty for interpretation and selected speeches containing the following:

- Regional accent
- Figures
- Acronyms
- Cultural References
- Proper nouns
- High speed
- Complex grammar and syntax
- Figures of speech
- Interruptions

Currently, AI interpretation uses three steps in a so-called cascade method: speech-to-text (STT), machine translation (MT) and text-to-speech (TTS). The aforementioned parameters can have an impact on some or all of the steps. So, we consider that these parameters are suitable for speech selection not only because they add difficulty to a real interpretation setting but also because they may be challenging for the cascade-method:

- Regional accent→ challenges STT
- Figures→ challenges STT and TTS
- Acronyms→ challenges STT and MT
- Cultural references → challenges MT

- Proper nouns→ challenges STT, MT and TTS
- High speed→ challenges STT and TTS
- Complex grammar and syntax→ challenges MT
- Figures of speech. → challenges MT
- Interruptions→ challenges STT

In a real-life setting, speakers are often interrupted by the chair. Thus, we selected almost half of the speeches with interruptions. Here, it is important to deal with both the transmission of the content and that of the interruption in order to be faithful to the communicative situation.

Interrupted by chair	Just interrupted	7	**44.4%**
	Engages in a parallel conversation	1	
Not interrupted by chair	-	10	**55.5%**

There is no such thing as an unaccented speech. Indeed, linguists prefer to talk about prestige and non-prestige variants, the first ones receiving the "standard" appellation and becoming the "norm". Although all speakers, even those who speak prestige dialects, have their accent, the exposure to certain accents and varieties is smaller than other more 'mainstream' varieties. Therefore, we decided to look for standard and regional accents of varying degrees.

Heavier accents (44.4%)	10	Arabic x1 Chinese x1 French x2 English x3 Russian x1 Spanish x2
Slight accents (55.5%)	8	Arabic x2 Chinese x2 French x1 Russian x2 Spanish x1

Speed of elocution is another variable that determines the difficulty of a speech. The ideal is a delivery of between 100-115 words per minute (wpm). In international organisations, the speed often increases by an additional 10-15 wpm. This parameter was taken into account as follows:

Slow-to-medium (5.6%)	1
Medium-to-fast (44.4%)	8
Very fast (50%)	9

Another parameter that can increase the difficulty of interpretation is dealing with sensitive content. Thus, some speeches are politically or emotionally charged whereas others are more technical. We also tried to find a middle ground in which political or emotive elements are present but are not the focus of the speech. Finally, we selected some 'simpler' speeches from a thematic point of view that do not have sensitive or highly technical elements and are more neutral in content.

Very politically/emotionally charged (22.2%)	4
Some political/emotional parts (33.3)	6
Very technical (16.7%)	3
Neutral content (27.8%)	5

Taking into account content, there are other aspects previously raised that can make a speech denser:

Figures	Many	13	77.8%
	Some	1	
	None	4	22.2%
Figures of speech	Many	6	38.9%
	Some	1	
	None	11	61.1%
Acronyms	Many	9	61.1%
	Some	2	
	None	7	38.9%
Complex grammar and syntax	Yes, many	6	72.2%
	Some	7	
	None	5	27.8%
Cultural References	Yes, many	13	83.3%
	Some	2	
	None	3	16.7%
Proper nouns	Yes	8	44.4%
	None	10	55.6%

In addition, we looked for other interesting and not so frequent elements that may also have an impact on interpretation. This includes speakers with speech difficulties (lisp, enunciation issues), cultural expressions or references and religious formulae.

We identified some very good speakers who make use of good rhetoric, with powerful emphatic pauses, and play with volume and intonation to highlight important keywords or ideas. In addition, we selected some speakers with a nervous or shaky voice, speakers who stumble over words or whose delivery is interspersed with voice clearing or coughing.

Finally, we wanted to cover the different types of sessions in WHA, therefore, we selected speeches from the plenary as well as from committees, to be as representative as possible. This is the resulting ratio:

Plenary (66.7%)	12
Committee A (22.2%)	4
Committee B (11.1%)	2

Having considered all those aspects, the following table summarizes the selected speeches:

LANGUAGE	SPEECH	TIMESTAMP	TOTAL TIME	SPEECH AVERAGE
ARABIC (16.65%)	Iraq (Plenary 3)	26:02-28:48 (2 min. 46 s.)	8 min 23 s.	2 min. 47 s.
	Oman (Committee A9)	02:22:51-02:26:05 (3 min 14 s.)		
	Yemen (Committee A10)	12:25:14:48 (2 min. 23 s.)		
CHINESE (16.65%)	China (plenary 5)	02:04:05-02:06:28 (2 min 23 s)	8 min. 6 s.	2 min. 42 s.
	China (plenary 3)	01:22:25-01:25:53 (3 min 28 s.)		
	China (Committee A13)	31:53-34:09 (2 min 15)		
FRENCH (16.65%)	RDC (Plenary 6)	35:54-38:12 (2 min. 18 s.)	7 min. 40 s.	2 min. 34 s.
	Haiti (Plenary 4)	02:06:03-02:09:24 (3 min. 21)		
	Senegal (Committee A13)	17:02-19:04 (2 min. 02 s.)		
ENGLISH (16.65%)	Malawi (Plenary 4)	01:17:24-01:20:46 (3 min 22 s.)	8 min. 9 s.	2 min. 43 s.
	Australia (Plenary 2)	02:47:57-02:50:57 (2 min.)		
	Bangladesh (Plenary 3)	2:32:33-2:35:20 (2 min 47 s.)		
RUSSIAN (16.65%)	Belarus (Plenary 3)	02:41:48-02:45:05 (3 min. 17 s.)	9 min. 5 s.	3 min 1 s.
	Russia (Plenary 3)	2:22:32-2:25:04 (2 min 32 s.)		
	Uzbekistan (Plenary 3)	02:12:17-02:15:33 (3 min 16 s.)		
SPANISH (16.65%)	Colombia (Committee B8)	03:09:45-03:13:12 (3 min 21 s.)	8 min. 20 s.	2 min. 46 s.
	Spain (Committee B8)	2:45:13-2:46:58 (1 min 45 s.)		
	Venezuela (Plenary 4)	09:52-13:06 (3 min 14 s.)		

We made sure all the languages had a similar amount of time, as shown in the table. The total average duration per speech is 2 min 46 s. For the test to be as realistic as possible, we kept the introduction of the speaker by the chair (in all cases in English) so as to test the “reflexes” of AI to code-switching and changing languages, but we did not add that to the time of English, despite needing interpretation, because we did not consider it to be a full speech.

ANNEX 2 – Breakdown of the selected speeches

SPEECH	Interrup.	Thick accent	Fig	Acron	Refer.	Prop. names	Fig speech	Complex gr/synt.	Speed	Emotionally or politically charged	OTHER ASPECTS
Iraq	-	-	Yes	No	Yes	No	Yes	Some	Medium	Yes	Cultural references/expressions Coughing, clearing voice…
Oman	Yes	-	Some	No	Yes	No	No	No	Fast	No (Technical)	
Yemen	-	Yes	Yes	Yes	Yes	No	No	Yes	Medium	No	Country speaking on behalf of a group of countries
China pl5	-	-	Yes	No	Yes	No	Yes	Yes	Medium	Some	Cultural references/expressions
China a13	Yes	-	Yes	Yes	Yes	No	No	Some	Fast	Some	
China pl3	-	Yes	Yes	No	No	Yes	Yes	Yes	Medium	Yes	Very good speaker
RDC	-	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Fast 172	Some	
Haiti	Yes	-	No	Yes	Some	No	Yes	Some	Fast 147	Some	Very good speaker
Senegal	-	Yes	Yes	Yes	No	No	No	No	Fast 176	No (Technical)	Nervous speaker, particularly challenging enunciation/accent
Malawi	Yes	Yes	Yes	Yes	Yes	Yes	No	No	Med 130	No	Speech stumbling
Australia	-	Yes	No	No	Yes	No	No	Some	Fast 143	Some	Speech stumbling
Bangladesh	-	Yes	Yes	Yes	Yes	Yes	No	Some	Fast 146	No	Lisp. Good speaker Cultural references/expressions
Belarus	Yes	-	Yes	Yes	No	No	No	No	Slow 113	No	Slowest speech
Russia	-	-	Yes	No	Yes	Yes	No	Some	Med 122	Some	Very good speaker
Uzbekistan	Yes	Yes	Yes	Some	Yes	Yes	No	Some	Med 123	no	
Colombia	Yes	-	No	No	Yes	Yes	Yes	Yes	Med 136	Yes	Speech stumbling Parallel interactions
Spain	-	Yes	No	Yes	Some	No	No	No	Fast 185	Some	Fastest speech
Venezuela	Yes	Yes	Yes	Some	Yes	Yes	Some	Yes	Fast 166	Yes	Very good speaker

ANNEX 3 – Quality assessment criteria for AI interpretation

	PARAMETERS	SCORING			
		UNACCEPTABLE (1)	WEAK (2)	SATISFACTORY (3)	STRONG (4)
CONTENT (A)	Coherence (a)	Message lacks logic (poor processing and analysis of content). The rendition does not make sense at all or is incomprehensible. Unsatisfactory intratextual coherence	Message lacks logic in large parts (unsatisfactory processing and analysis of content). Some parts are incomprehensible. Intratextual coherence is lost at times.	Logic is mostly conveyed (satisfactory processing and analysis of content). Negligible and irrelevant logic issues. Intratextual coherence mainly kept.	Logical message carefully processed and analysed. Intratextual coherence well kept.
	Completeness and accuracy (b)	Omissions, additions or changes in meaning twist the message. The rendition is not faithful at all. Significant hallucinations.	Some omissions, additions or changes of meaning alter the message. The rendition varies with faithful parts and inaccurate parts. Insignificant hallucinations.	Irrelevant omission or additions. Slight changes in meaning. The rendition is mostly accurate and complete. No hallucinations.	Faithful, complete and accurate message. Omissions or additions are useful. No hallucinations.
	Meaning (c)	Primary meaning as well as secondary ideas are mostly distorted (large meaning shifts), implicit ideas are lost. Incorrect division/merge of sentences and/or phrases.	Primary meaning is frequently distorted. Secondary ideas and some implicit ideas are lost. Incorrect division/merge of sentences and/or phrases.	Few negligible meaning shifts, but primary meaning is mostly conveyed. Practically all secondary ideas are kept. Some implicit ideas are lost. No division/merge of sentences and/or phrases.	Primary meaning is conveyed and all secondary ideas are kept. Implicit meaning is also successfully conveyed. No division/merge of sentences and/or phrases.
	Reputational damage/risks (d)	The content of the message threatens or ridicules the identity, values, image of the organization, member state or speaker	The content of the message involves inaccuracies regarding the identity, values, image of the organization, member state or speaker	The content of the message does not involve major inaccuracies regarding the identity, values, image of the organization, member state or speaker	The content of the message fully aligns with the identity, values, image of the organization, member state or speaker
EXPRESSION AND FORM (B)	Use of language (a)	Mostly incorrect grammar and syntax.	Often incorrect grammar and syntax	Mostly correct grammar and syntax	Wholly correct grammar and syntax
	Idiomatic expression (b)	Poor handling of structural differences in SL&TL, poor collocations, literal translation or unnatural expression.	Weak handling of structural differences in SL&TL, weak collocations, some instances of literal translation or unnatural expression.	Satisfactory handling of structural differences in SL&TL, effective collocations, some instances of literal translation or unnatural expression.	Good handling of structural differences in SL&TL, rich collocations, idiomatic expression.

	Word choice and terminology (c)	Mostly inappropriate terms, poor vocabulary, frequent use of ineffective calques. Institutional/cultural and other references incorrectly translated and/or addressed. Wholly unsuitable register for the context. Offensive word choice.	Often inappropriate terms, average vocabulary, some calques Some institutional/cultural and other references correctly translated and/or addressed but others aren't. Sudden inadequate changes of register.	Mostly appropriate terms, good vocabulary but with some calques. Some institutional/cultural and other references correctly translated and/or addressed but others aren't. Acceptable choice of register for the given context.	Expert terminology used, rich vocabulary, rare instances of calques, if any. All the institutional/cultural and other references incorrectly translated and/or addressed Good register.
DELIVERY (C)	**Prosody (a)**	Doesn't flow smoothly. Unsteady volume, poor voice clarity, unnatural-sounding intonation. Largely incorrect oral punctuation.	Flow could be smoother. Not the best volume/voice projection, largely unnatural-sounding intonation. Some incorrect oral punctuation.	Generally pleasant and smooth but some parts are burdensome to follow. Acceptable volume/voice projection, natural cadence despite some rare instances of unnatural intonation. Mostly correct oral punctuation.	Very smooth and pleasant. Close to a human natural delivery. Voice-projection close to that of a human, natural-sounding intonation. Correct oral punctuation.
	Vocal output (b)	Generally unpleasant, robotic sound. Fragmented rhythm with long silences, sudden changes in pacing, bad vocalization, major pronunciation errors (syllable stress, tones…)	Largely unpleasant and robotic. Large parts with rhythm and uncomfortable silences, generally good pacing but suddenly raced at times, some vocalization issues, some pronunciation errors (syllable stress, tones…)	Rhythm and pace are mostly easy to follow. Acceptable vocalization and mostly good pronunciation (syllable stress, tones…)	Rhythm and pace are easy to follow. Human-like vocalization and pronunciation (syllable stress, tones…)
	Communicability (c)	Emotions and/or purpose of the message is not at all rendered. Key information or pieces highlighted by the speaker (with volume, intonation, stressed…) are not noticeable or not at all rendered	Some emotional charge is lost. Purpose is not successfully rendered. Some parts highlighted by the speaker (with volume, intonation, stressed…) are not noticeable or not at all rendered and others are (inconsistencies)	Some emotional charge is conveyed. Purpose is mostly conveyed. Key information or pieces highlighted (with volume, intonation, stressed…) are noticeable and rendered	Emotional charge is conveyed. Purpose is fully conveyed. Key information or pieces highlighted (with volume, intonation, stressed…) are noticeable and rendered

ANNEX 4 – Raw overall data and summary of reputational risks

INTO / FROM	ARABIC		CHINESE		ENGLISH		FRENCH		RUSSIAN		SPANISH	
ARABIC												
Oman			44%	1 RR	50%	1 RR	63%	1 RR	56%	2 RR	38%	1 RR
Yemen			27%	1 RR	70%	2 RR	47%	1 RR	37%	1 RR	30%	6 RR
Iraq (Multilingual)			50%	2 RR	67%	2 RR	46%	5 RR	62%	3 RR	46%	2 RR
CHINESE												
China – 1	64%	3 RR			63%	3 RR	52%	6 RR	36%	3 RR	62%	1 RR
China – 2	38%	1 RR			58%	3 RR	38%	4 RR	50%	1 RR	38%	1 RR
China (Multilingual)	38%	1 RR			42%	2 RR	36%	3 RR	19%	1 RR	39%	2 RR
ENGLISH												
Australia	58%	1 RR	53%	1 RR			83%	1 RR	63%	1 RR	70%	2 RR
Bangladesh	38%	5 RR	43%	1 RR			60%	3 RR	60%	3 RR	43%	2 RR
Malawi (Multilingual)	36%	3 RR	57%	4 RR			43%	5 RR	53%	4 RR	50%	2 RR
FRENCH												
Haiti	48%	1 RR	58%	2 RR	69%	1 RR			60%	1 RR	64%	2 RR
Senegal	27%	3 RR	7%	5 RR	31%	5 RR			30%	1 RR	30%	2 RR
DRC (Multilingual)	5%	9 RR	15%	2 RR	28%	6 RR			22%	4 RR	39%	3 RR
RUSSIAN												
Russia	70%	1 RR	50%	2 RR	71%	1 RR	47%	2 RR			20%	2 RR
Uzbekistan	20%	2 RR	13%	5 RR	16%	4 RR	15%	1 RR			22%	3 RR
Belarus (Multilingual)	40%	4 RR	53%	1 RR	61%	7 RR	42%	4 RR			50%	3 RR
SPANISH												
Spain	62%	2 RR	73%	2 RR	57%	3 RR	69%	2 RR	64%	2 RR		
Venezuela	54%	2 RR	27%	3 RR	45%	2 RR	57%	1 RR	43%	1 RR		
Colombia (Multilingual)	38%	3 RR	30%	3 RR	33%	4 RR	54%	1 RR	17%	4 RR		
FINAL ASSESSMENT	**42%**	**41 RR**	**40%**	**35 RR**	**51%**	**46 RR**	**50%**	**40 RR**	**45%**	**32 RR**	**45%**	**34 RR**

ANNEX 4a - Average scores per speech as well as average per language

INTO FROM	ARABIC	CHINESE	ENGLISH	FRENCH	RUSSIAN	SPANISH	**FINAL AVERAGE**
Oman		44%	50%	63%	56%	38%	
Yemen		27%	70%	47%	37%	30%	
Iraq (Multilingual)		50%	67%	46%	62%	46%	
ARABIC		40%	62%	52%	52%	38%	**49%**
China – 1	64%		63%	52%	36%	62%	
China – 2	38%		58%	38%	50%	38%	
China (Multilingual)	38%		42%	36%	19%	39%	
CHINESE	47%		54%	42%	35%	46%	**45%**
Australia	58%	53%		83%	63%	70%	
Bangladesh	38%	43%		60%	60%	43%	
Malawi (Multilingual)	36%	57%		43%	53%	50%	
ENGLISH	44%	51%		62%	59%	54%	**54%**
Haiti	48%	58%	69%		60%	64%	
Senegal	27%	7%	31%		30%	30%	
DRC (Multilingual)	5%	15%	28%		22%	39%	
FRENCH	27%	27%	43%		37%	44%	**36%**
Russia	70%	50%	71%	47%		20%	
Uzbekistan	20%	13%	16%	15%		22%	
Belarus (Multilingual)	40%	53%	61%	42%		50%	
RUSSIAN	43%	39%	49%	35%		41%	**41%**
Spain	62%	73%	57%	69%	64%		
Venezuela	54%	27%	45%	57%	43%		
Colombia (Multilingual)	38%	30%	33%	54%	17%		
SPANISH	51%	43%	45%	60%	41%		**48%**
FINAL AVERAGE	**41%**	**40%**	**51%**	**50%**	**45%**	**45%**	**46%**

ANNEX 4b - Average scores for languages

INTO / FROM	ARABIC	CHINESE	ENGLISH	FRENCH	RUSSIAN	SPANISH	**FINAL AVERAGE**
ARABIC		40%	62%	52%	52%	38%	**49%**
CHINESE	47%		54%	42%	35%	46%	**45%**
ENGLISH	44%	51%		62%	59%	54%	**54%**
FRENCH	27%	27%	43%		37%	44%	**36%**
RUSSIAN	43%	39%	49%	35%		41%	**41%**
SPANISH	51%	43%	45%	60%	41%		**48%**
FINAL AVERAGE	**41%**	**40%**	**51%**	**50%**	**45%**	**45%**	**46%**

GRADES:	**Fail (0%-74%)**	**Pass (75%-100%)**

ANNEX 5 – Average scores in the bilingual series (when source languages were preset)

INTO FROM	ARABIC	CHINESE	ENGLISH	FRENCH	RUSSIAN	SPANISH	FINAL AVERAGE
Oman		44%	50%	63%	56%	38%	
Yemen		27%	70%	47%	37%	30%	
ARABIC		36%	60%	55%	47%	34%	**46%**
China – 1	64%		63%	52%	36%	62%	
China – 2	38%		58%	38%	50%	38%	
CHINESE	51%		61%	45%	43%	50%	**50%**
Australia	58%	53%		83%	63%	70%	
Bangladesh	38%	43%		60%	60%	43%	
ENGLISH	50%	48%		72%	62%	57%	**58%**
Haiti	48%	58%	69%		60%	64%	
Senegal	27%	7%	31%		30%	30%	
FRENCH	38%	33%	50%		45%	47%	**43%**
Russia	70%	50%	71%	47%		20%	
Uzbekistan	20%	13%	16%	15%		22%	
RUSSIAN	45%	32%	44%	31%		36%	**38%**
Spain	62%	73%	57%	69%	64%		
Venezuela	54%	27%	45%	57%	43%		
SPANISH	58%	50%	51%	63%	54%		**55%**
FINAL AVERAGE	**48%**	**40%**	**53%**	**53%**	**50%**	**45%**	**48%**

GRADES:	**Fail (0%-74%)**	**Pass (75%-100%)**

ANNEX 6 – Average scores in the multilingual series (when the source languages were not specified)

INTO FROM	ARABIC	CHINESE	ENGLISH	FRENCH	RUSSIAN	SPANISH	**FINAL AVERAGE**
ARABIC (Multilingual)		50%	67%	46%	62%	46%	**54%**
CHINESE (Multilingual)*	38%		42%	36%	19%	39%	**35%**
ENGLISH (Multilingual)	36%	57%		43%	53%	50%	**48%**
FRENCH (Multilingual)	5%	15%	28%		22%	39%	**22%**
RUSSIAN (Multilingual)	40%	53%	61%	42%		50%	**49%**
SPANISH (Multilingual)	38%	30%	33%	54%	17%		**34%**
FINAL AVERAGE	**31%**	**41%**	**46%**	**44%**	**35%**	**45%**	**40%**

*The output of the interpretation from the Chinese speech in the multilingual series was a blind assessment, that is an assessment of the output without checking the source. The output was therefore assessed in all languages by non-Chinese speakers, to check the pure comprehensibility of the output.