Automatic Speech Recognition and Understanding in Air Traffic Control – How much Information is lost over the Atlantic Ocean?

Hartmut Helmke
Institute of Flight Guidance
German Aerospace Center (DLR)
Braunschweig, Germany
hartmut.helmke@dlr.de

Stephen S. Clarke
Systems Modeling & Optimization
NASA Ames Research Center
Moffett Field, California, US
stephen.s.clarke@nasa.gov

Oliver Ohneiser
Institute of Flight Guidance
German Aerospace Center (DLR)
Braunschweig, Germany
oliver.ohneiser@dlr.de

Matthias Kleinert
Institute of Flight Guidance
German Aerospace Center (DLR)
Braunschweig, Germany
matthias.kleinert@dlr.de

Shruthi Shetty
Institute of Flight Guidance
German Aerospace Center (DLR)
Braunschweig, Germany
shruthi.shetty@dlr.de

Krishna Kalyanam
Aviation Systems Division
NASA Ames Research Center
Moffett Field, California, US
krishna.kalyanam@nasa.gov

***Abstract*—We compare two ontologies – from US and Europe – for interpretation of semantics of controller and pilot radio telephony transcriptions. Overall, more than 2000 ground control transmissions are evaluated from both sides of the Atlantic. Half of the transmissions, from the US side, are transformed into semantic concepts with the European ontology. These concepts contain the information, which the ontology claims as relevant. The concepts are transformed back into sequences of words, which are then transformed back to semantic concepts by the US ontology, i.e., the US implementation of what information is relevant. Comparing the US concepts generated through this process against the original US concepts defines the European loss. The process is then executed the other way around on the second half of transmissions stemming from the European side, resulting in the US loss. The European command type loss is currently 2.6% absolute, whereas the US loss is 1.2%, i.e., the extraction rate of commands decreases by these losses over the Atlantic Ocean. In addition to loss comparison, the main contribution of this work is that the generalized approach enables comparison and evaluation of different ontologies and especially different implementations of the ontologies.**

Keywords—Automatic Speech Understanding, Ontology, Air Traffic Control

I. INTRODUCTION

A. Background

Automatic speech recognition (ASR) aims to transcribe a sequence of words from a speech or audio sample. During the last decade many ASR applications in the air traffic management (ATM) domain have been developed around the globe.

The challenge, however, is to perform speech understanding for which different ontologies have been proposed. An ontology is an abstraction from the word level into higher level air traffic control (ATC) concepts such as callsigns or command types like descend, vectors, taxi instructions, etc. On the one hand, abstractions simplify subsequent tasks, while on the other hand information can be lost with each abstraction step. This paper presents an approach to quantify the information, which is lost when different ontologies from the US and Europe are applied on ASR outputs from controller utterances across Europe and the US.

Researchers in the US and Europe have independently developed ATC language ontologies for different control areas. Which ontology is the *best* one? MITRE and DLR have already compared their ontologies in 2023 [1] with emphasis on approach control. The result was that the choice of the ontology depends on the application.

B. Air Traffic Control Communication Contents

This paper compares a European and a US language ontology for the apron and ground area. Is it possible to use the US ontology without changes for ground movements at a European airport and vice versa? Of course not, but what is lost? This paper, in addition to presenting a qualitative analysis of two ontologies, also develops a framework, which enables the comparison of different ontologies and their implementations.

Over 1000 ground control transmissions, represented as word sequences (transcriptions) from three US top-30 passenger volume airports were collected as well as over 1000 transcriptions from three different apron positions from a European top-10 airport. All transcriptions also come with prepared semantic annotations into their respective ontology, i.e., US transcriptions come with semantic annotations created via the US ontology and the same is true for the European data. The US transcriptions are then transformed by the European ontology and respectively the same is done with the European transcriptions via the US ontology.[1] This means that we generate semantic annotations, i.e., the transformation into well-defined relevant information items, for all roughly 2000 transcriptions using the opposite ontologies.

For example, the original US transcription "*Delta Seven thirty-five runway three four right de-ice pad, taxi via Yankee.*" results in the original US annotation "DAL735 (**deice**:runway_34R) Y". However, the same US transcription would result in the European annotation "DAL735 TAXI TO

[1] The authors were asked by DASC program committee to enable an anonymous review. Therefore, we call the ontologies the *European ontology* and the *US ontology*, although currently no comparable standardization activity neither on word level nor on semantic level exists in the US. The European ontology is a handover from the two EUROCAE working groups WG41 and WG126 for standardization.

RW34R, DAL735 TAXI VIA Y". The information about the de-icing action is lost in the European ontology.

All annotations are then transformed back into word sequences, which often results in transcriptions that deviate from the ground truth, i.e., the original transcript. Next the original and generated transcripts are used as input into the ontology matching the origin of the original transcript, i.e., if the original transcript comes from the US both transcripts are transformed to semantic annotations via the US ontology and vice versa for the EU. The difference of the semantic annotations between the original and the generated transcript is the trans-Atlantic annotation loss.

This loss can be seen on word level, i.e., the transcriptions after transformation and the original transcriptions differ. Recalling the above example, transforming back into the original US annotation to a word sequence results in "*delta seven three five* ***deice*** *runway three four right taxi via yankee*". However, when transforming back the European annotation to a word sequence, it results in "*delta seven three five taxi to runway three four right taxi via yankee*". The way the flight number of the callsign was uttered is different from the original US transcription. Furthermore, we see the loss of "deice" on European side stemming from the annotation loss, i.e., missing or modified ATC concepts in the annotation. Also note the differences in how numbers are represented in the callsign. Although the meaning has not changed, "*seven thirty-five*" got converted to "*seven three five*" when converted by both ontologies. In our paper, we focus on the annotation loss. This loss can easily be seen when comparing the original US annotation to the lossy US annotation after trans-Atlantic transformations – based on the generated transcripts on European side – being only "DAL735 (runway_34R) Y" without "deice".

C. Paper Structure

Section II presents related work on ATC ontologies, before we explain the two ontologies for comparison on word and semantic level in section III. Section IV describes our performed experiments to get quantitative results, which are presented in section V. Section VI discusses the results and gives an outlook on whether harmonization is a good idea and how to do it. Section VII presents our conclusions.

II. Related Work

Researchers all over the world have independently developed rules to note down the meaning of ATC radiotelephony utterances. This section describes the differences in existing ontologies for extracting of semantic contents from ATC utterance transcriptions. All these ontologies set rules for the annotation of defined ATC-specific terminology with potentially assigned values and for entities dependent on airspace, airport, or aircraft. In other words, they care about aircraft callsigns, types of clearances / commands / intents, values from a range / set / slot fillers like altitudes, waypoints, runway names, taxiway names, etc. but encapsulate these rules in different syntax and terms as shown above.

A. Existing Ontologies for Air Traffic Control

As the flight callsign – usually consisting of an airline designator and a flight number with digits and letters – is one of the most important semantic content for further analyzing the remainder of an utterance transcript. All explored ontologies define the callsign as an ATC concept. To achieve high recognition rates in extracting callsigns, a list of currently available callsigns should be used as contextual knowledge, e.g., stemming from flight plans [2], Mode-S data [3], or surveillance data such as automatic dependent surveillance-broadcast (ADS-B) [4], [5]. Furthermore, the clipping, skipping, and misrecognition of digits or numbers in the callsign recognition process has been considered [6]. The callsign similarity rules including format definitions defined by Eurocontrol also help to acquire the correct callsign from a transcription [7]. The above cited works just concentrate on the ontology element *callsign* and follow the rules of the International Civil Aviation Organization (ICAO) for the annotation. Callsign examples are *DLH7HT, AAL1099, SIA807* and *OENKF*.

For all other elements of the explored ontologies, the rules and syntax are much more heterogeneous than for the callsign. The extraction of different types of utterance content can be called *detection of controlling intent* [8], *event detection* [9], or *command recognition* [10]. ATC-related classes for semantic extraction are for example called *level change communication* with a specific event Code *Av* [11] or simply *climb*, *descend*, etc. [8]. The natural language processing (NLP) task of *slot filling* then extracts the required instruction elements for the above command types like altitude and speed values, but also airline designators and flight numbers for callsigns [8]. Also, slot filling combined with the NLP task *intent classification* has also been used for digitization of ATCo commands by extraction into an ontology using deep learning [12].

An NLP-inspired ontology for conversation understanding between controllers and pilots was suggested by researchers from Pakistan [13]. They use opening and closing extensible markup language (XML) tags such as <*ORGANIZATION*> for airlines, <*LOCATION*> for airports, or <*DATE*> for times in the named entity recognition task. A similar XML-tagging is used in a German ontology that proposes a semantic layer with, e.g., <callsign>, <airline>, <flightnumber>, <commands> <command = "type">, etc. and a concept layer with, e.g., "*DLH3RK REDUCE 240*" [14]. One ontology implementation from Singapore includes 40 distinct categories for commands of air traffic controllers (ATCos) and responses by pilots such as *DESCEND, REDUCE SPEED, HEADING* [15]. Another ontology implementation of MITRE from the US distinguishes command types, qualifiers, and parameters and intensely compares the different existing command types and their usage in various downstream applications to a European ontology [16]. Lastly, the ICAO phraseology [17] or the controller-pilot data link communications (CPDLC) data message format [17] can be seen as further ontologies to communicate ATC radiotelephony content.

There exist additional ontologies in a broader ATM context that deal with named entities of aviation infrastructure or flight safety messages that can as well be part of radiotelephony communication. The NASA ATM ontology has been compared with an ontology that was derived from the "ATM Information Reference Model" [18]. Another ontology implementation from Iran concentrates on extracting the concepts within flight safety messages for ATM [19].

The above-described ontologies have a lot of commonalities, but – depending on the technological maturity and the use case – differ especially for the representation of ATC command contents. Although there have been many ontologies created for ATC and ATM, our paper focuses on specific ones that have been developed that pertain to data collected at the European and US airports of interest.

B. Ontologies for Comparison in this Paper

The European ontology for ATC utterance annotations used for the analysis in this paper assumes that each utterance consists of a callsign and one or more commands. Each command can consist of type, value, unit, qualifier, and conditions. In total, there are more than one hundred different command types defined. This basic scheme of an ontology has been agreed between more than 20 European ATM stakeholders [20]. To better distinguish between transcripts and semantic annotations, the main elements of the command element annotations use majuscules (see also Table I in section III.B). A French implementation heavily relies on this European ontology and adapts it to its own needs where required [21].

Building off previous work, the US ontology used for the analysis in this paper has been adapted to include commonly used phraseology found in the FAA Order JO 7110 [22] and phraseology found in operational data that extended beyond published standards. The operational data analyzed included ground and apron commands from the three major US airports of interest.

III. US AND EUROPEAN ONTOLOGIES

Now we will discuss the similarities and differences between the two ontologies used in this paper. According to [1], communications can be considered in terms of four levels of a computer interaction model consisting of *lexical*, *syntactical*, *semantic*, and *conceptual* levels [23].

The **lexical level** deals with words and distinguishes between synonyms – different words which have the same meaning in different contexts or utterances. These words are also the building blocks for ATC radiotelephony transmissions. This level of an ontology specifies the elements that may appear in a transmission. The vocabulary consists of general-purpose words such as *taxi*, *give*, *way*, *pushback* as well as names, such as those for airline callsigns, location identifiers etc. Ideally, these terms are static and can be defined as part of an official vocabulary list or dictionary in the ontology.

The **syntactical level** deals with grammar and distinguishes between similar meaning phrases that are worded differently. For example, the phrases "*taxi two seven left via yankee and november*" and "*taxi yankee november to two seven left*" are different on the syntactical level because they use different words, but have the same meaning, just describing a taxi clearance given the same taxi route and final destination.

The **semantic level** deals with the meaning of individual transmissions. We include information not explicitly spoken but implied in the transmission. Both example phrases from the syntactical level may be mapped to an agreed form such as the European form "*TAXI VIA Y N, TAXI 27L*" or the US form "*(dest:runway_27L) Y N*".

The **conceptual level** deals with a higher level of understanding that goes beyond the semantic level, which often depends on the application itself. An ideal ontology is independent of its software implementation and will support a wide range of downstream and end-user applications [23]. In this paper, we primarily address the lexical and semantic level described above since each application of the ontology can differ.

A. Ontology for Transcriptions on the Word Level

This subsection presents the US and European ontology at the lexical level, i.e., at the word level. "*Delta Sixteen thirteen ground roger, runway zero niner right, taxi via Alpha, Echo, hold short Hotel, and at Two West you're gonna give way to inbound traffic it's a three nineteen.*" shows the word level representation. The European representation of the same sentence is "***d****elta* ***s****ixteen thirteen ground roger runway zero* ***nine*** *right, taxi via* ***alf****a echo, hold short* ***h****otel, and at* ***t****wo* ***w****est* ***you are*** *gonna give way to inbound traffic* ***it is A*** *three nineteen".*

- The European ontology for transcription has no upper-case letters with one exception that a letter is pronounced like in the airline designator KLM, ILS, or the first letter of the alphabet in "A three nineteen".
- The European version has no punctuation marks like commas, dots, or quotation marks. However, the US uses punctuation to represent intonation in speech *"taxi via echo hotel" (EH)* versus *"taxi via echo, hotel" (E H).*
- The European version avoids the apostrophe. It uses "*you are*" instead of "*you're.*"
- Europe uses the ICAO/NATO alphabet letters *alfa*, *juliett*, and *whiskey*, whereas US uses *Alpha*, *Juliet*, and *Whisky*.
- The name of the airline and the first number of a callsign starts with an upper-case letter in the US version.
- The European ontology transcribe "tree" and "niner" as "three" and "nine" whereas the US transcribes exactly what is said.

The transcription of a spoken sentence is also called the "gold transcription", if it is the hypothetical output of a perfect speech-to-text engine. The US version already contains semantic information at the word level as some words are written with upper case letters depending on context. Also, the use of punctuation marks can give better understanding, when creating the semantic interpretation in subsequent steps.

B. Ontology Considerations for Semantic Annotations

The example in table I shows a transmission containing a callsign and a taxi clearance. The semantic interpretation is displayed in both a simplified human-readable format and a machine-readable JSON format with blue highlighted keywords for the succeeding values. Note the difference between US (SL_{US}) and European semantic interpretations (SL_{EU}). Here, SL_{EU} uses the same syntax as in [20] by condensing the callsign to a single string, as it would appear in an electronic flight strip. On the other hand, SL_{US} consists of several elements that reconstruct the final callsign. This is to enable partially captured or malformed callsigns.

TABLE I. SEMANTIC INTERPRETATION OF SIMPLE TAXI CLEARANCE

delta thirty seven ten taxi yankee	
SL$_{US}$	DAL3710 Y
[{"callsign": "DAL3710", "3ltr": "DAL", "callsign_number": "3710"}, {"taxiway": "Y"}]	
SL$_{EU}$	DAL3710 (TAXI VIA) Y
{"csgn": "DAL3710", "type": "TAXI", "sndT": "VIA", "valu": "Y"}	

In the US ontology, each command is split into sequential components. The component could be a callsign, taxiway, spot, the *at* keyword, or an action. The full list of current command types can be found in table XV in section V. In addition to each command type, an attribute is used to give context and meaning to the component. When converting between syntactic and semantic levels, the US format splits a callsign into

different parts in the JSON format, (i) three letter code (3ltr), (ii) the letter and number part, and (iii) the callsign modifier, which can contain the wake category "heavy" or "super tug". Actions are split into (i) action type, (ii) attribute, and (iii) designator.

The European ontology presses the whole callsign into one single string – even in the JSON format. A transmission consists of one or multiple instruction(s) in both ontologies. In the European ontology, an instruction requires a command and a callsign, which is *NO_CALLSIGN*, if no callsign is said or a callsign in not extracted. The elements of the command are best explained by Fig. 1 taken from [24].

Fig. 1. Block diagram of SL_{EU} ontology, optional elements are in orange.

A *Command* concept always has a *Type* attribute, which could be *NO_CONCEPT*. A Type can be split into a main type and a subtype like "*TAXI VIA*" in Table I. One or more values may follow. If a *Value* is available, the optional attributes *Unit* and *Qualifier* can follow. The optional *Condition* component can be present for any Type. Speaker is *ATCo* or *PILOT*. If missing, *ATCo* is assumed. Reason can be *REQUEST*, *REPORTING*, or *READBACK*. The last one is assumed, when the reason is missing.

C. Comparison of Example Annotations

TABLE II. EXAMPLE OF TRANSMISSION WITH 3 INSTRUCTIONS

Delta Thirty-nine Salt Lake ground, Alpha, Echo, hold short abeam Hotel.	
***SL**US*	DAL39 A E (hold:H)
[{"Callsign": "DAL39", …}, {"taxiway": "A"}, {"taxiway": "E"}, {"action": "hold", "attribute": "taxiway", "designator": "H"}]	
***SL**EU*	DAL3739 STATION SL_GND, DAL3739 (TAXI VIA) A E, DAL3739 HOLD_SHORT H WHEN ABEAM
[… {"csgn": "DAL3739", "type": "STATION", "valu": "SL_GND"}, {"csgn": "DAL3739", "type": "TAXI", "sndT": "VIA", "valu": "A", "vare": "E"}, {"csgn": "DAL3739", "type": "HOLD_SHORT", "valu": "H", "qual": "ABEAM"}]	

Table II shows a transmission transcribed in US format resulting in a *taxi* and a *hold* instruction in US annotation format and three instructions – *STATION*, *TAXI*, *HOLD_SHORT* – in European annotation format. In the following examples we use grey boxes for the transcriptions, which are sometimes in US and other times in EU format. An obvious difference is the extracted callsign. SL_{EU} uses context information for callsign interpretation. Hence, "37" is added to the flight number, because *DAL3739* is the only callsign of this airline currently under control of this ATCo. SL_{US} contains only the spoken words of the callsign; it is the task of downstream applications to figure out the associated callsign. SL_{EU} adds the *TAXI VIA* command type, although not spoken. SL_{US} only models the taxiway values A and E. It contains the information that A and E are taxiways, whereas SL_{EU} models this implicitly. "*TAXI VIA*" is always followed by a taxiway. The "hold" action of SL_{US} corresponds to the *HOLD_SHORT* type of SL_{EU}. SL_{EU} lists the callsign in all instructions although only spoken once. This eases the representation of transmissions containing instructions for different callsigns. The qualifier *ABEAM* and the word sequence "*Salt Lake ground*" are not modeled by SL_{US}.

We omit the callsign information and the machine-readable JSON format in the following examples concentrating on the relevant information to explain differences. Table III shows an example with qualifier (*RIGHT*) and condition (*WHEN AT N7*). The gray boxes sometimes show the US transcription format and other times the European format.

TABLE III. GIVE_WAY INSTRUCTION

at november seven give way to lufthansa seven three from right to left	
***SL**US*	(at:N7, giveway:"lufthansa seven three from right to left")
***SL**EU*	GIVE_WAY (DLH B737) RIGHT WHEN AT N7

SL_{US} uses the action *giveway*, SL_{EU} the *GIVE_WAY* type. SL_{EU} fully tries to interpret the command, whereas SL_{US} passes the string "*lufthansa seven three from right to left*" to preserve the natural language description. In the same way SL_{EU} interprets "seven three" as a "boeing 737", whereas SL_{US} keeps 73 without further interpretation. If interpretation is already done by the ontology, this eases the task of downstream applications, but may lead to a loss of information. SL_{EU} just contains that the *lufthansa* comes from the right side, but misses the information that it passes from *right to left*.

The tables IV to VII show further examples for transfer to downstream applications.

TABLE IV. REPORT_MISCELLANEOUS

good morning who is going to forty three	
***SL**US*	(question: "Who's going to forty three")
***SL**EU*	GREETING; REPORT_MISCELLANEOUS

No information is lost in SL_{US} for the downstream applications for the example shown in table IV, except the *GREETING*, which is only modeled by SL_{EU}.

TABLE V. CONTINUE TAXI AND REPORT_MISCELLANEOUS

continue the echo sierra route kilo to the ramp what is your entry	
***SL**US*	ES K (dest:ramp) (question:entry)
***SL**EU*	(CONTINUE TAXI); (TAXI VIA) ES K; (TAXI TO RAMP); REPORT_MISCELLANEOUS

SL_{US} interprets the words "*what is your entry*" as a question for the entry in table V. SL_{EU} marks this part as a general question to report something. SL_{US} ignores the *continue* of the taxi instruction, which can be important if an aircraft stopped during the taxi process.

TABLE VI. TRAFFIC INFORMATION AND STANDBY

Six forty six you'll see a Southwest there off your right, uh they're pulling in to gate thirty eight, you might have to wait just a second till they pull in, I'm not sure if that, uh that'll work.	
***SL**US*	646 (standby: "you might have to wait just a second till they pull in, I'm not sure if that, uh that'll work")
***SL**EU*	VXP646 CALL_YOU_BACK; VXP646 (INFORMATION TRAFFIC)

In table VI SL_{US} models the action as *standby* and SL_{EU} uses the type *CALL_YOU_BACK*. The text "*Southwest there off your right, uh they're pulling in to gate thirty eight*" is lost in both ontologies. SL_{US} adds a long text to the *standby* action. SL_{EU} ignores the 40 words just interpreting them as traffic information, but adds *VPX* meaning *avelo* to the callsign.

TABLE VII. REQUEST VERSUS NO_CONCEPT

yeah, just if I need to move you, just uhm I'll let you know ahead of time.	
SL_{US}	(request: "if I need to move you, just I'll let you know ahead of time")
SL_{EU}	NO_CONCEPT

Table VII shows an example, which results in NO_CONCEPT in SL_{EU}. Further examples will follow, when presenting the quantitative results.

IV. PERFORMED EXPERIMENTS

A total of 1016 ATC transcriptions represented as word sequences from apron controllers of the three US airports are modeled by the European ontology. The output abstractions, i.e. the annotations, are then transformed back into a sequence of words. The resulting sequence of words and the original transcription are often different. Both are used as input into the US ontology resulting in annotations. The difference between the two resulting annotations is the annotation loss of the European ontology. The process has already been described with a colored example in section I.B.

A. Transformation Process for Loss Calculation

We formalize this process now: The US transcriptions are transformed into the European transcription format by the transformation function $t_{US\rightarrow EU}(ws)$, which maps a sequence of words *ws* from the US into the European transcription format both described in subsection III.A. $t_{EU\rightarrow US}(ws)$ is the corresponding function for mapping from the European to the US transcription format. Note that $t_{EU\rightarrow US}(t_{US\rightarrow EU}(ws))$ is not always equal to *ws*, e.g. , the information of capital letters is lost.

The annotation function $a_{US}(ws)$ transforms a given word sequence *ws* into a US semantic level representation. $a_{EU}(ws, c)$ transforms a given word sequence *ws* into the European semantic level representations. $a_{EU}(ws, c)$ relies on context information *c*, which consists of a dynamic and a static part. The dynamic part includes the list of available callsigns, which are currently under control of the speaking ATCo. Conversely, $a_{US}(ws)$ does not currently utilize any context information. Usually, context information is provided by external processes, e.g., a controller assistance system to perform *assistant based speech recognition* (ABSR) [10]. The word sequence "*two zero twenty three salt lake ground"* can then be transformed into, e.g., DAL2023 or AAL2023 depending on the callsigns actively under control of the ATCo, when the word sequence was generated. ABSR uses this context information to correct wrong outputs of the speech to text process.

For our paper, the list of callsigns for US airports was not directly available. Therefore, the context information was artificially generated by post-processing. We assumed that the callsigns used at least once in the utterances of each of the corresponding airports were in context. Normally, the currently active automatic terminal information service (ATIS) information and the air pressure also belongs to the dynamic context information.

Additionally, the static information consists of all runways, taxi points, taxiways, etc. Therefore, the word sequence "*ground, runway three four right"* results in the named entities *GND* and *TRW34*, because the static context contains that the word sequence "*ground"* should be mapped to *GND* and the word sequence "*runway three four right"* corresponds to *TRW34* for the *TAXI TO* command type. Table VIII shows the number of named entities used for the different US airports and European apron positions.

TABLE VIII. NUMBER OF NAMED ENTITIES

	US Airports			***European Positions***		
	A1	***A2***	***A3***	***P1***	***P2***	***P3***
Aircraft Types	13	30	0	25	25	25
Frequencies	5	3	1	12	12	12
Runways	6	10	8			
Taxi Points	28	30	12	256	257	255
Taxiways	47	36	28	72	74	72

Both the functions $a_{US}(ws)$ and $a_{EU}(ws, c)$ could be calculated either manually or using some automatic process, which converts the natural language into the semantic representation. In this paper, $a_{US}(ws)$ is calculated manually and $a_{EU}(ws, c)$ automatically. The automatic process for $a_{EU}(ws, c)$ is described in [25] .

The inverse function $s_{EU}(sr, c)$ transforms a semantic representation *sr* in European format to a possible word sequence by also using context information *c*. For example, "*DAL735 TAXI VIA A"* could be mapped to "*delta seven three five taxi via alfa*". It is important to note that this function is not unique. The output for the above input could also be "*delta thirty five use alfa*". For the analysis in this paper, we always use an identical unique word sequence version. We do not want to evaluate the extraction capacities on the other side of the Atlantic, but just the lost information resulting from the ontology. To summarize: Given a word sequence *ws* spoken by a US ATCo, we perform the steps with examples shown in table IX:

TABLE IX. STEPS TO DETERMINE LOSS FROM US TO EUROPE

	Step	***Who***	***How***
1	$a_{US}(ws) \rightarrow sr^{Org}{}_{US}$	US	M
Delta Forty-nine de-ice pad taxi then via Alpha. → DAL49 (deice) A			
2	$t_{US\rightarrow EU}(ws) \rightarrow ws_{EU}$	EU	A
Delta Forty-nine de-ice pad taxi then via Alpha.→ delta forty nine deice pad taxi then via alfa			
3	$a_{EU}(ws_{EU}, c) \rightarrow sr_{EU}$	EU	A
delta forty nine deice pad taxi then via alfa (context: only current flight from DAL has number 1149) → DAL1149 TAXI VIA A			
4	$s_{EU}(sr_{EU}, c) \rightarrow ws^{*}{}_{EU}$	EU	A
DAL1149 TAXI VIA A → delta one one four nine taxi via alfa			
5	$t_{EU\rightarrow US}(ws^{*}{}_{EU}) \rightarrow ws^{*}{}_{US}$	US	A
delta one one four nine taxi via alfa → Delta Eleven Forty-nine taxi via Alpha.			
6	$a_{US}(ws^{*}{}_{US}) \rightarrow sr^{*}{}_{US}$	US	M
Delta Eleven Forty-nine taxi via Alpha. → DAL1149 A			
7	$loss(sr^{Org}{}_{US,},sr^{*}{}_{US})$	US	A
loss(DAL49 (deice) A, DAL1149 A) = 1 deletion			

Column *Step* lists the transformation functions. Column *Who* refers to the trans-Atlantic region to perform an action in the process. Column *How* indicates if it is an automatic *(A)* or manual process *(M)*. The function *loss* calculates the difference between both semantic representations, i.e., semantic loss on the European side. The difference is calculated using the Levenshtein distance [26] between both representations. With *loss* we distinguish between command recognition rate, command recognition error rate, and command rejection rate,

whose sum could be more than 100%, e.g., when one command is in the ground truth, but three commands are extracted. More details are provided in [16]. Each row below a numbered row provides an example for a succeeding transformation. The last row shows that "*deice*" contributes as deletion to the loss. The complete callsign *DAL1149* with "11" has been interpreted given contextual knowledge in the European ontology implementation while the US ontology leaves this functionality for later applications. This is information *gain* across the Atlantic, it is not calculated as loss.

The same formal process is repeated by using 1037 utterances from a European airport with three different apron positions as input. Annotations are created using the US ontology, which is then transformed back into a sequence of words. Both sequences of words are inputted into the European ontology and compared, resulting in the annotation loss of US ontology. The European transcription "*hello lufthansa four zero two pushback is approved area three*" is annotated as "*DLH402 (push:approved)*" and backward transcribed as "*lufthansa four zero two pushback approved*". The greeting and the "*area three*" information are lost.

Fig. 2 illustrates a more general version of this process, which can be applied to any two arbitrary ontologies that represent text. The word sequence *ws* is interpreted by the two ontologies to produce the semantic representations sr_1 and sr_2. considering a transformation by t by step 2 of the words. The second semantic representation is then reconstructed as text *ws** and a transformation of the words by step 6, which is semantically interpreted by the first ontology as sr_1*. Finally, the loss between the two outputs sr_1 and sr_1* is determined.

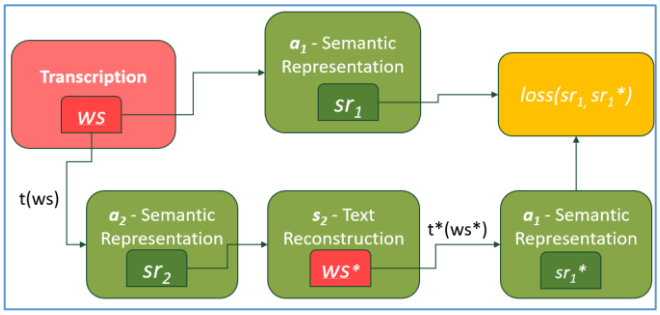

Fig. 2. Steps to determine the Trans-Atlantic Loss

B. Pre-Evaluation Of US Reconstruction

A first pre-evaluation was performed with a sub-set of 280 commands resulting from 120 of the 1037 European transmissions to reveal some systematic differences that might be better ignored for the final evaluation. The results of the first pre-evaluation show the US ontology performance *US-Pre1* in table X.

TABLE X. INITIAL SEMANTIC LEVEL PERFORMANCE CONSIDERING GREETINGS/FAREWELLS IN [%]

	Command Level				Callsign Level	
	Recogn	***Error***	***Precision***	***Recall***	***Recogn***	***Error***
US-Pre1	58.6	13.2	81.6	66.4	99.2	0.0

First, we see that we have no loss for the callsigns. In one transmission, a callsign is corrected by the ATCo. This correction is lost, i.e., the wrong callsign is not contained in step 4. Only 58.6% of the commands are retained from EU to US and back. Analysis shows that SL_{US} ignores the *GREETING* and *FAREWELL* types as the US ontology does not see any value in those types. We also noticed that the *PUSHBACK* and *CONTINUE TAXI* are very often not correctly modeled by the US ontology. "*pushback area one*" or "*pushback facing east*" both result into the reconstructed transmission "pushback". "*continue lima november six november*" results e.g. into "*taxi via lima november six november*". The application of the US ontology [12] does not need these command types. To make a more meaningful and fair comparison, we decided to ignore these types from the results presented in section V.

C. Pre-Evaluation Of European Reconstruction

Comparable to the previous subsection a pre-evaluation of the steps shown in table IX was performed. Loss values of nearly 20% were observed in step 7, which mostly result from step 3 and 4, i.e. the automatic transformation of a US utterance into the European ontology ($a_{EU}(ws_{EU}, c) \rightarrow sr_{EU}$) and sometimes from the backward-transformation into a word sequence. For example, "*Kilo's now current*" was interpreted as a taxi clearance via taxiway Kilo and not as ATIS information. "*push runway one five left*", was interpreted as a pushback clearance and a taxi clearance to runway 15L and not as just an information that a taxi clearance to 15L can be expected after pushback. "*point sixty five on top of the bridge"* was modeled as "CONTACT_FREQUENCY 123.650 WHEN AT BRIDGE". The reconstructed output, however, was "contact one two three decimal six five zero", which loses the information that the frequency should not be changed now, but later, when the aircraft is on the bridge. This is, however, not a problem of the ontology, but of step 4 the re-construction of a possible transmission. Automatic semantic interpretations were, therefore, manually corrected.

D. Finding Loss due to Annotation and Reconstruction

The previous subsection shows that the reason for a high loss value might not be the modeling deficiencies of an ontology, but automatic annotation or the reconstruction process. Another reason could be step 1 or step 6, i.e. the manual annotation on US side. We, therefore, performed another experiment just ignoring the manual steps on US side. US transmissions were manually annotated (gold annotations) in the EU ontology (step 3 in table IX) $\mathbf{sr^M_{EU}}$. For the gold annotations word sequences $\mathbf{ws^*_{EU}}$ were automatically generated (step 4 in table IX). These word sequences ($\mathbf{ws^*_{EU}}$) are than automatically transformed back into the EU ontology and compared against the manually created gold annotations $\mathbf{sr^M_{EU}}$.

TABLE XI. RECONSTRUCTION PERFORMANCE OF US COMMANDS IN [%]

Command Type	Airport #1		Airport #2		Airport #3	
	Prec.	**Rec.**	Prec.	Rec.	**Prec.**	**Rec.**
AFFIRM	*100*	*100*	*100*	*100*	***100***	***100***
CONT. TAXI	***100***	***100***	***100***	***100***	*100*	*100*
CROSS	*0*	*N/A*	*N/A*	*N/A*	***73***	***83***
FOLLOW	***70***	***100***	*N/A*	*N/A*	*43*	*100*
GIVE_WAY	***75***	***100***	***68***	***93***	*100*	*100*
HOLD_SHORT	***95***	***95***	*90*	*100*	***75***	***100***
INFORM ATIS	***100***	***95***	*100*	*100*	*100*	*67*
PUSHBACK	*100*	*100*	*N/A*	*N/A*	***100***	***100***
REPORT_MIS	***100***	***100***	***100***	***100***	***100***	***100***
STATION	***100***	***99***	***100***	***100***	***100***	***100***
TAXI TO	***99***	***98***	***98***	***99***	***96***	***97***
TAXI VIA	***98***	***99***	***99***	***99***	***98***	***100***
TURN	***95***	***90***	***94***	***100***	***74***	***100***

Table XI contains recall and precision of different command types of the European ontology representing US transmissions. Only types occurring at least 20 times at least at one US airport are presented (**bold font,** if occurring at least 20 times for that airport). Red font indicates that the recall is below 95%. Cells with *N/A* show that certain commands never happened at some airports. Analyzing the TURN command, we see that there are improvements to be made in the ontology. One known limitation is that if two turns occur in the same instruction, then the second turn would be lost. The words of the *CROSS* command are automatically generated, but the automatic extraction of *CROSS* with a condition has problems and "*taxi via charlie cross runway one nine left*" often results into a "*TAXI TO RW_19L*". For "Airport#1" all CROSS commands contain a condition. The current implementation of *GIVE_WAY* only expects aircraft types like "*airbus three twenty*", but not spot names or taxiway names.

Table XII presents the results, when the semantic interpretation is directly performed on the original US transmission transcriptions. Of course, performance drops down as much more word variations are used in real ops room transmissions. We highlighted with yellow background, when we observed a big difference between the two tables XI and XII. The biggest difference is observed for *REPORT MISCEL LANENOUS*, which should not surprise, because no detailed concept exists for this command type. Therefore, the implementation can only implement them as heuristics.

TABLE XII. EXTRACTION PERFORMANCE OF US COMMANDS IN [%]

Command Type	Airport #1		Airport #2		Airport #3	
	Prec.	**Rec.**	**Prec.**	**Rec.**	**Prec.**	**Rec.**
AFFIRM	*100*	*100*	*100*	*100*	***100***	***97***
CONT. TAXI	***100***	***100***	***100***	***96***	*75*	*100*
CROSS	*0*	*N/A*	*N/A*	*N/A*	***46***	***75***
FOLLOW	***35***	***100***	*N/A*	*N/A*	*14*	*50*
GIVE_WAY	***57***	***95***	***67***	***93***	*100*	*57*
HOLD_SHORT	***90***	***92***	*90*	*100*	***75***	***100***
INFORM ATIS	***100***	***58***	*80*	*33*	*100*	*67*
PUSHBACK	*50*	*75*	*N/A*	*N/A*	***100***	***94***
REPORT_MIS	***74***	***56***	***100***	***79***	***65***	***33***
STATION	***100***	***100***	***99***	***99***	***97***	***100***
TAXI TO	***95***	***91***	***92***	***97***	***95***	***80***
TAXI VIA	***89***	***93***	***92***	***92***	***94***	***92***
TURN	***71***	***75***	***88***	***88***	***62***	***57***

In summary, we can now explicitly see a difference in loss between the two ontologies, but also whether the difference results from the automatic annotation or automatic transmission generation.

E. Correcting the Loss of Automatic Reconstruction

We then performed a second pre-evaluation with all 1037 European transmissions to reveal further systematic differences in the rest of the data that we might want to exclude from the full evaluation. The results and some calculation adjustments of the second pre-evaluation are shown in table XIII.

When evaluating all manually, perfectly annotated transmissions, one would expect a recognition of 100%. However, when automatically extracting the semantics of the European transcriptions (*EU-Pre2a*), i.e., step 6 in table IX, the command recognition is just 97.7%. The software is automatically not able to parse all manually created annotations correctly, so that we end up with an accuracy slightly below 100%. We define the result of *EU-Pre2a* as the *Ground Truth* for annotations of European transmissions in table XIII.

The row *US-Pre2a* shows the US ontology performance for annotating European transmissions. However, the performance of *US-Pre2a* is most probably better than the calculated result as the *Ground Truth* is 2.3% (100% - 97.7%) away from the correct result. Hence, we calculate correction values (row *Corr. Value*) as one minus the *Ground Truth* and the *Ground Truth* itself for columns showing errors. The last row *US-Pre2b* shows the corrected results, i.e., *US-Pre2a* plus or minus *Corr. Value* with some rounding differences. This correction mechanism is used for all further evaluations.

TABLE XIII. SEMANTIC LEVEL PERFORMANCE WITHOUT GREETINGS/FAREWELLS INCLUDING CORRECTION VALUES IN [%]

	Command Level				Callsign Level	
	Recogn	***Error***	***Precision***	***Recall***	***Recogn***	***Error***
Ground Truth / EU-Pre2a	97.7	1.1	98.9	98.3	99.2	0.3
US-Pre2a	68.8	15.4	81.7	79.7	99.2	0.0
Corr. Value	2.3	1.1	1.1	1.7	0.8	0.3
US-Pre2b	71.0	14.4	82.8	81.4	100	0.0

V. RESULTS

This section presents the semantic level performance of the US and European ontology, dives deeper into reasons for different performances on command type level and and extracted attributes. Table XIV shows the comparison of each recognized command in ground truth sr^{Org}_{EU} compared with the reconstruction sr^{*}_{EU}. Column *Total* counts how often each *Command Type* was observed within the 1037 European transmissions.

TABLE XIV. US SEMANTIC LEVEL PERFORMANCE PER COMMAND TYPE FOR EUROPEAN TRANSMISSIONS (PRECISION AND RECALL IN [%])

Command Type	Total	Only Type		All Except Callsign	
		Prec.	**Recall**	**Prec.**	**Recall**
CALL_YOU_BACK	*49*	*100*	*100*	*100*	*100*
CONTACT	*153*	*97*	*99*	*97*	*99*
CONTACT_FREQ	*274*	*100*	*100*	*99*	*100*
CONTINUE TAXI	~~*188*~~	*N/A*	*N/A*	*N/A*	*N/A*
CORRECTION	*20*	*100*	*5*	*50*	*5*
FAREWELL	~~*65*~~	*N/A*	*N/A*	*N/A*	*N/A*
GIVE_WAY	*120*	*100*	*98*	*80*	*96*
GREETING	~~*235*~~	*N/A*	*N/A*	*N/A*	*N/A*
HOLD_POSITION	*46*	*100*	*100*	*100*	*100*
HOLD_SHORT	*330*	*99*	*100*	*96*	*100*
INFORM TRAFFIC	*9*	*100*	*89*	*100*	*89*
NO_CONCEPT	*13*	*N/A*	*N/A*	*N/A*	*N/A*
PUSHBACK	~~*122*~~	*N/A*	*N/A*	*N/A*	*N/A*
TAXI TO	*267*	*100*	*99*	*100*	*99*
TAXI VIA	*509*	*99*	*98*	*96*	*98*
TURN	*66*	*98*	*100*	*97*	*100*
Summary	***1964***	***99.4***	***98.2***	***96.9***	***98.2***
Loss		***1.2***		***2.5***	

The row *Summary* shows the sum of *Total* command types and the average for the other columns. We only concentrate on command types occurring at least five times, i.e., we ignored 13 of the 1964 commands, e.g., command type *DISREGARD*. We already explained, why we excluded

GREETING/*FARWELL*, whose *Total* is omitted from the evaluation. The next two columns show precision and recall for each command type, when we only check whether the correct type is extracted *(Only Type)*. The last two columns show precision and recall, when the whole command except the callsign is considered, i.e., there is an error, if the type, value(s), qualifiers, or the conditions are wrongly extracted *(All Except Callsign)*. If only the callsign is wrongly extracted from the reconstructed transmission, we assume it as correct. Lastly, the *Loss* is calculated by calculating the inverse F1-score between the total precision and recall.

Keep in mind that the extraction is performed automatically and thus is not perfect. We, therefore, performed the automatic extraction from the original transmission and then subtracted these errors from the errors of the reconstruction. However, only 18 errors were observed from the original ones: 7 false positives and 11 false negatives. The row *Summary* considers the corrected Total, true positives etc. of each command type. Rows with *N/A* are ignored in the calculation of Precision and Recall. 2466 is the sum of all command types if rows with N/A are not omitted.

The cells marked in red from table XIV are explained in more detail: the US ontology does not consider the *CORRECTION* type at all. Therefore, we have very low recall. It is not 0% due to the *Corr. Value* explained by table XIII. For *GIVE_WAY*, the qualifiers *LEFT* and *RIGHT* were often lost. Most of the *HOLD_SHORT* problems result from errors during automatic extraction, e.g., sometimes "continue up to november one" was reconstructed as "taxi via november one". As already explained in subsection IV.B, we excluded the command types *GREETING*, *FAREWELL*, *CONTINUE TAXI* and *PUSHBACK* from the calculation of the last two rows.

TABLE XV. EUROPEAN SEMANTIC LEVEL PERFORMANCE PER COMMAND TYPE FOR US TRANSMISSIONS (PRECISION AND RECALL IN [%])

Command Type	Total	Only Type		All Attributes	
		Prec.	Recall	Prec.	Recall
Action: ATIS	*36*	*100*	*100*	*100*	*100*
Action: Contact (1)	*74*	*94.9*	*100*	*94.9*	*100*
Action: Cross	*36*	*97.2*	*97.2*	*91.7*	*91.7*
Action: De-Ice (2)	*16*	*100*	*100*	*87.5*	*87.5*
Action: Destination	*491*	*99.2*	*98.6*	*98.2*	*97.6*
Action: Exit	*6*	*100*	*100*	*100*	*100*
Action: Expect (3)	*50*	*100*	*98*	*100*	*98*
Action: Follow (4)	*28*	*100*	*96.4*	*85.2*	*82.1*
Action: Give To You (5)	*5*	*-*	*0*	*-*	*0*
Action: Give Way (4)	*119*	*96.7*	*99.2*	*86.1*	*88.2*
Action: Hold	*82*	*100*	*100*	*95.1*	*95.1*
Action: Push	*57*	*100*	*100*	*100*	*100*
Action: Question (6)	~~*170*~~	*N/A*	*N/A*	*N/A*	*N/A*
Action: Request (6)	~~*80*~~	*N/A*	*N/A*	*N/A*	*N/A*
Action: Squawk	*7*	*100*	*100*	*100*	*100*
Action: Standby	~~*18*~~	*N/A*	*N/A*	*N/A*	*N/A*
Action: Turn	*125*	*92*	*82.4*	*91.1*	*81.6*
Action: Who (7)	~~*7*~~	*N/A*	*N/A*	*N/A*	*N/A*
At Bridge	*17*	*100*	*100*	*100*	*100*
At Taxiway	*98*	*85.2*	*76.5*	*84.1*	*75.5*
Callsign (8)	*924*	*99.3*	*99.1*	*97.3*	*97.1*
Spot (7)	*13*	*100*	*7.7*	*100*	*7.7*
Taxiway	*949*	*98.8*	*97.1*	*97.5*	*95.8*
Summary	***3133***	***98.3***	***96.5***	***96.2***	***94.5***
Loss		***2.6***		***4.6***	

Table XV shows the same comparison as table XIV, but this time the 1016 US transcriptions are interpreted by the EU ontology, transformed to word sequences and then the new word sequences are automatically interpreted by the US ontology as described by the steps in table IX. Thus, it compares each recognized command in ground truth sr^{Org}_{US} with the reconstruction sr^{*}_{US}. Note that here, *Callsign* is considered a command type since it is extracted individually from a sequence instead of captured for each command like in the European ontology. Now we explain some differences in detail and use the numbers in column one as references:

(1) Monitor is a variation of *Contact*. Instead of calling another frequency the pilot is asked to switch to another frequency and wait for instruction. Although seldom used, there is no European representation for these types of commands.

(2) Europe has no special action for deice, but uses the taxi point de-ice pad as the end of the taxi route, which sometimes results in lost information.

(3) SL_{EU} only has *EXPECT RUNWAY*, but no *EXPECT TAXIWAY*.

(4) FOLLOW and GIVE_WAY were sometimes mixed and slot positions were not modeled. "*from left to right*" and "*give way to two or more aircraft*" were not modeled.

(5) The *GiveToYou* command is not modeled. It is given to pilots, when they approach an intersection and another aircraft in their vicinity should give way to them. Although not commonly used, we chose to represent this concept so that it can be captured by downstream applications.

(6) As shown in table IV, *question* and *request* are not modeled by the European ontology and SL_{US} only models natural language information so they are excluded from evaluation.

(7) European ATCos use callsigns or avoid the callsign, when it is clear who is addressed. The transmissions "*spot twenty two continue turning left*" has no callsign, because the ATCo does not know who is at spot 22. In some airports, pilots will approach and wait at spot numbers before being issued taxi commands. Once situated at a spot, ATCos will refer to the spot number to give commands like *"continue turning left"* to begin taxiing to a runway. Quite often they follow up with the question *"who are you"* to identify the callsign or signal the pilot to turn on their transponder so that the ATCo knows the callsign of the aircraft at the spot.

(8) SL_{EU} does not consider callsign modifiers like *heavy*.

VI. DISCUSSION AND NEXT STEPS

Throughout this research, both sides of the Atlantic not only learned about the information lost, but also strategies to build more resilient ontologies, when representing ATC instructions. Although the focus of this research was to compare the ontologies with respect to modeling the semantics of ATC transmissions, there are as well lessons to be learned in the extraction and processing of data. For example, the European side uses context information about callsigns outside the utterance itself. This process is invaluable when interpreting utterances that are often noisy or clipped while recorded. Although not currently implemented during the US extraction, this would be a powerful tool to implement in the future.

We also have a better understanding of how each ontology could improve in specific areas. The *giveway* instruction is a

good example how both ontologies and especially the underlying extraction algorithm could benefit from each other. The US ontology was able to model the transmissions in table XVI. The original version of the European ontology expects an aircraft type as the second value part.

TABLE XVI. GIVE_WAY-INSTRUCTION WITH SLOT INFORMATION

give way to delta at spot forty two	
SL_{US}	(at:spot_42,giveway:DAL)
SL_{EU}	GIVE_WAY (DAL none) *new*: GIVE_WAY (DAL SPOT42)

The US ontology benefits from capturing left/right information as already shown in table III. While the European ontology was able to represent complex commands, the US ontology defaulted to using strings to capture the natural language. These strings, although capturing some of the nuances in how controllers give commands, is less interpretable by automation systems.

Some differences in the ontologies result from different operational behavior on both sides of the Atlantic, e.g. using spot numbers instead of callsigns. These types of operations, although specific to certain airports, are still important to capture. This example also highlights that airport-specific operations exist. Most US airports do not use spots in this way. European ATCos prefer the word "*decimal*" whereas US ATCos prefer "*point*" in frequency changes like "*contact ground one two one decimal eight five*" versus "*contact ground point eight five*". This example also shows that US ATCos prefer shorter clearances. The European shortening of the frequency change clearance might be "*two one eight five*", but the part before the word *point* is mostly said.

The ten digits from "zero" to "nine" cover 42% of all words observed in the European data set. In the US data set, the same digits comprise only 29% of all spoken words [16]. ATCos and pilots are not limited to the ten digits, as recommended by ICAO [12]. US ATCos and pilots prefer the other group-form digit words such as "ten", "twenty", "thirteen", "fourteen" etc. Looking at callsigns, we see that in the US, ATCos tend to combine digits in pairs whereas in Europe ATCos tend to say each digit e.g., "*united twelve thirty four*" versus "*united one two three four.*" When these additional numbers are summed up together with "zero" through "nine", then numerical words comprise 40% of all words spoken in the US, compared to 42% in Europe [16].

In the US data set, ATCos typically use more general English descriptors when giving instructions to pilots. Words like "*behind*" are used to cover a range of scenarios. After deeper analysis though, we are able to distil words like *behind* into more widely used "*give way*" or "*follow*". In the example "*alpine eleven behind UPS proceed to the ramp*" the European reconstruction appears as "*alpine air one one give way to UPS taxi to the ramp*." Both have the same meaning, but the ATCo decided to use a generic English term "*behind*" to describe *give way* in this case. Lastly, we discuss how to calculate and interpret loss of information. We also found two general types of loss between each ontology. (i) Loss, which occurs, when a concept is only represented by one ontology although concepts exist in both, e.g. "WHEN AT BRIDGE". (ii) Loss, if a concept does not exist in both datasets, and therefore is lost by one ontology. For example, deicing pads are only used at airports in cold weather and would be lost if developing an ontology using data from airports in tropical environments.

We plan to broaden the research in several directions. The current focus on this paper was on ground transmissions given the scope of previous work. Future work should expand to Tower, TRACON, and Center transmissions to find more differences between the operations in the US and Europe. Additionally, after seeing the specific nuances of each airport, there is more to be learned from other major airports and facilities throughout the US and Europe.

The main contribution of this paper is the presented 7-step-approach, which enables us to compare US and European ontologies and shows the first steps for harmonization. Using gold transmissions *ws* (manual transcriptions) and gold semantic interpretations sr^{Or}, this approach is able to calculate the loss between any two ontologies and helps identifying areas of improvement. Forms of improvement include (i) unifying similar terminology, such as *exit* and *vacate* and (ii) finding gaps in either ontology where the reconstruction loss is bigger than 0%. Additionally, this method could be extended to other ontologies and not just within the aviation domain. If there are any two representations of the same data, this method could be used to find differences and enable harmonization. On the other hand, the 7-step approach could also enable the comparison of different implementations of semantic extraction algorithms. This can be done by using gold transcriptions *ws*, gold semantic interpretations sr^{Or}, and their extracted semantic interpretations sr^{*} from an automatic extraction algorithm. The calculated difference – the loss – between sr^{Or} and sr^{*}can show where improvements can be made to the extraction method. If no gold transcriptions are available, they can be automatically generated from artificial semantic interpretations by function *s* (step 4 in table IX**).** Noise can be randomly added to the semantic interpretations to test the robustness of the automatic extraction function. Although the focus of this paper is on the gold semantic interpretations, future work could include these comparisons of extraction algorithms across the Atlantic.

In the future, we plan to explore different automatic extraction methods, Currently, the European extraction uses a complex set of rules to capture and convert transmissions into their semantic interpretations. Building on previous work, we want to understand how to improve this extraction process using both deep learning [12], rules [25] and hybrid workflows. Provided this workflow for calculating extraction loss, we will be able to better improve the results from these methods. Another potential extraction method is also the use of Large Language Models (LLMs). Given the recent acceleration and advancements in LLMs, they could provide better extraction accuracy and potentially reduce the data required for training through prompt tuning.

VII. CONCLUSIONS

One contribution of this paper is a step towards harmonizing European and US ontologies for ATC applications. The main contribution, however, is that this approach is generalizable and enables researchers to compare different ontologies and evaluate their implementations. It also enables the comparison of rule-based extraction implementations of an ontology against machine-learning based implementations and other methods. Furthermore, our approach does not raise any data privacy issue as no voice recordings need to be exchanged. Furthermore, no manual annotations, i.e. semantic

interpretations, need to be exchanged. They are not even needed, if the aim is just to compare the differences in ontology implementations.

Lastly, we have been able to calculate the transatlantic loss between US and Europe. We show that the European loss on operational data is 2.6% on the command level and 4.6% considering all information, whereas the US loss on lab data is 1.2% on the command level and 2.5% considering attribute information. These losses arose from (i) differences in phraseology used between the US and Europe as well as airport-specific operations and (ii) concepts that exist on both sides of the Atlantic, but were not captured by their respective implementation due to a focus in downstream application. Analyzing these losses and individual command types allows us to find specific improvements to make for each ontology. The acceptable good values require, however, some pre-processing steps especially adapting the different syntactical representations on word level. Otherwise losses of bigger than 90% would have been measured.

ACKNOWLEDGMENT

We thank the retired ATCos and colleagues for their tedious work in transcribing and annotating the analyzed transmissions.

REFERENCES

[1] H. Helmke, O. Ohneiser, M. Kleinert, S. Chen, H.D. Kopald, and R.M. Tarakan, “Transatlantic Approaches for Automatic Speech Understanding in Air Traffic Management,” 15th USA/Europe Air Traffic Management Research and Development Seminar (ATM2023), Savannah, GA, USA, 5–9 June 2023.

[2] R. García, J. Albarrán, A. Fabio, F. Celorrio, C. Pinto de Oliveira, and C. Bárcena, “Automatic Flight Callsign Identification on a Controller Working Position: Real-Time Simulation and Analysis of Operational Recordings,” Aerospace 2023, 10, 433. doi: 10.3390/aerospace10050433.

[3] J. Zuluaga-Gomez, K. Veselý, A. Blatt, P. Motlicek, D. Klakow, A. Tart, I. Szöke, A. Prasad, S. Sarfjoo, P. Kolčárek, et al., “Automatic Call Sign Detection: Matching Air Surveillance Data with Air Traffic Spoken Communications,” Proceedings 2020, 59, 14. doi: 10.3390/proceedings2020059014.

[4] M.S. Kasttet, A. Lyhyaoui, D. Zbakh, A. Aramja, and A. Kachkari, “Toward Effective Aircraft Call Sign Detection Using Fuzzy String-Matching between ASR and ADS-B Data,” Aerospace 2024, 11, 32. doi: 10.3390/aerospace11010032.

[5] J.M. Madhathil, N.N. Khanh, L. Seounghoon, L.T. Tuan, T.A. Dung, and T.H. Dat, ”Automatic Speech Recognition and its Contextual Enhancement for Singapore ATC Voice Communication,” SESAR Innovation Days, Rome, Italy, 12–15 Nov 2024.

[6] A. Blatt, M. Kocour, K. Veselý, I. Szöke, and D. Klakow, “Call-Sign Recognition and Understanding for Noisy Air-Traffic Transcripts Using Surveillance Information,” ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 8357–8361, doi: 10.1109/ICASSP43922.2022.9746301.

[7] “Call Sign Similarity Rules,” EUROCONTROL, 2022. [Online]. Available: https://www.eurocontrol.int/sites/default/files/2022-06/eurocontrol-call-sign-similarity-rules.pdf

[8] Y. Lin, “Spoken Instruction Understanding in Air Traffic Control: Challenge, Technique, and Application,” Aerospace 2021, 8, 65. doi: 10.3390/aerospace8030065.

[9] J.M. Cordero, M. Dorado, and J.M. de Pablo, “Automated speech recognition in ATC environment,” In Proceedings of the 2nd International Conference on Application and Theory of Automation in Command and Control Systems (ATACCS '12). IRIT Press, Toulouse, France, 2012, pp. 46–53.

[10] H. Helmke, J. Rataj, T. Mühlhausen, O. Ohneiser, H. Ehr, M. Kleinert, Y. Oualil, and M. Schulder, “Assistant-Based Speech Recognition for ATM Applications,” 11th USA/ Europe Air Traffic Management Research and Development Seminar (ATM2015), Lisbon, Portugal, 2015.

[11] N.R. Ucles and J.M. Cordero, “Relationship between Workload and Duration of ATC Voice Communications,” 6th International Conference on Research in Air Transportation, Istanbul, Turkey, 26–30 May, 2014.

[12] H.A. Steinmetz, J. Tao, S.S. Clarke, and K. Kalyanam, ”A Natural Language Understanding Approach for Digitizing Aircraft Ground Taxi Instructions,” AIAA AVIATION FORUM AND ASCEND, Las Vegas, USA, 2024.

[13] D. Abdullah, H. Takahashi, and U. Lakhani, “Domain Specific Ontology Enhancing Communication Accuracy in Airport Operation,” 2019 IEEE 14th International Symposium on Autonomous Decentralized System (ISADS), Utrecht, Netherlands, 2019, pp. 1-5, doi: 10.1109/ISADS45777.2019.9155591.

[14] M. Schulder, J. O’Mahony, Y. Bakanouski, and D. Klakow, “ATC-ANNO: Semantic Annotation for Air Traffic Control with Assistive Auto-Annotation,” In Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 6375–6380, Marseille, France. European Language Resources Association, 2020.

[15] P.T. Dat, L.T. Tuan, J.M. Madhathil, and T.H. Dat, “Automatic Speech Recognition and Understanding Over Noisy Air Traffic Control VHF Channels in Singapore,” SESAR Innovation Days, Rome, Italy, 12–15 Nov 2024.

[16] S. Chen, H. Helmke, R.M. Tarakan, O. Ohneiser, H. Kopald, and M. Kleinert, “Effects of Language Ontology on Transatlantic Automatic Speech Understanding Research Collaboration in the Air Traffic Management Domain,” Aerospace 2023, 10, 526. doi: 10.3390/aerospace10060526.

[17] “Procedures for Air Navigation Services (PANS)—Air Traffic Management (Doc 4444),” International Civil Aviation Organization (ICAO): Montreal, QC, Canada, 2016.

[18] E. Gringinger, R.M. Keller, A. Vennesland, C.G. Schuetz, and B. Neumayr, “A Comparative Study of Two Complex Ontologies in Air Traffic Management,” 2019 IEEE/AIAA 38th Digital Avionics Systems Conference (DASC), San Diego, CA, USA, 2019, pp. 1-9, doi: 10.1109/DASC43569.2019.9081790.

[19] M. Yousefzadeh Aghdam, S.R. Kamel Tabbakh, S.J. Mahdavi Chabok, and M. kheyrabadi, “Ontology generation for flight safety messages in air traffic management,” J Big Data, 8, 61, 2021, doi: 10.1186/s40537-021-00449-3.

[20] H. Helmke, M. Slotty, M. Poiger, D.F. Herrer, O. Ohneiser, N. Vink, A. Cerna, P. Hartikainen, B. Josefsson, D. Langr, R. García Lasheras, G. Marin, O.G. Mevatne, S. Moos, M.N. Nilsson, and M.B. Pérez, ”Ontology for transcription of ATC speech commands of SESAR 2020 solution PJ.16-04,” IEEE/AIAA 37th Digital Avionics Systems Conference (DASC), London, United Kingdom, 2018.

[21] F. Betouret Candau, J. Carol, S. Martin, and T. Roques, “Scribe NLP: Unleashing the potential of ATC Voice Communication,” 2024 International Workshop on ATM/CNS (IWAC2024), Tokyo, Japan, 19–20 Nov, 2024.

[22] “FAA Order JO 7110.65BB – Air Traffic Control,” Federal Aviation Administration, Washington, D.C., 2025

[23] J.D. Foley and A. Van Dam, “Fundamentals of interactive computer graphics,” 1st edition, Reading, MA, USA: Addison-Wesley Publishing Company, 1982.

[24] M. Kleinert, S. Shetty, O. Ohneiser, H. Ehr, H. Ariliusson, T.S. Simiganoschi, A. Prasad, P. Motlicek, K. Vesely, K. Ondrej, P. Smrz, J. Harfmann, and C. Windisch, “Readback error detection by automatic speech recognition to increase ATM safety,” 14th USA/Europe Air Traffic Management Research and Development Seminar (ATM2021), Virtual Conference, 2021.

[25] M. Kleinert. H. Helmke, S. Shetty, O. Ohneise. H. Ehr, A. Prasad, P. Motlicek, and J. Harfmann, “Automated Interpretation of Air Traffic Control Communication: The Journey from Spoken Words to a Deeper Understanding of the Meaning,” IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA, 2021.

[26] V.I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” in: Soviet Physics -- Doklady 10.8, Feb. 1966.