22nd ITS World Congress, Bordeaux, France, 5–9 October 2015

Paper number ITS-2955

When Will ITS Speak Your Language?

Tamás Váradi[1*], Marko Tadić[2], András Gulyás[3], Mihai Niculescu[4]
1. Hungarian Academy of Sciences Research Institute of Linguistics, Hungary, 1068 Budapest, Benczúr u. 33, e-mail: varadi.tamas@nytud.mta.hu, phone: +36203441403
2. Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
3. University of Pécs, Faculty of Civil Engineering and Informatics, Hungary
4. ITS Romania, Romania

Abstract
The paper proposes a solution to the issue of language which is underrated in the present-day practice and policy of ITS. Drawing on best practices and mature technology in ITS and language technology respectively, an Internet based cross-border real-time traffic information system is envisaged in which road users will receive traffic information spoken in their native language through their journey across Europe. In addition, the service will bring to their car information relevant to their current position and destination. The combination of high quality machine translation and speech technology, coupled with the DATEX II ontology, promises a feasible and sustainable system, ready to be scaled up to a pan-European level. The system is amenable to deployment in a wide variety of settings including being embedded in the quickly spreading global navigation satellite systems.

KEYWORDS:
RTTIS, ITS, language technology.

1. The problem

In the modern age traffic is inherently international. It flows across national borders and as a consequence even within the territory of any single country a sizeable part of road users are foreign drivers. This is especially predominant in Europe, which consists of 45 countries using almost the same number of national languages. Within the European Union, national borders no longer constitute a barrier to the free flow of people and goods. However, the uncomfortable truth is that decades after the breaking down of national borders within the EU, a much more forceful and universal barriers remain: language barriers. With 24 official national languages within the European Union alone this presents a huge challenge for real-time traffic information services (RTTIS). Language barriers prevent road users from receiving pre-trip and, more crucially, on-trip real-time traffic information in their native language.

2. The state of the art

Real-time traffic-information services in Europe are characterized by fragmentation and isolation by language and even by country boundaries. National RTTIS centers provide information in the national language of the country with additional service (often on a limited scale) in a foreign language, typically English. It is obviously inadequate for drivers travelling abroad who find themselves isolated in a foreign language medium. What is more, it is also inadequate as a national service, i.e. in terms of reaching out to the drivers on the national road network because at any time a significant number of drivers are in transit across

the country and presumably a large part of them do not speak the local language. According to a survey performed in e.g. Austria, the share of transit vehicles was as high as 64.5% in freight transport alone in 2009.
Traditionally, road traffic information is provided over the radio and it is still the medium most preferred by drivers despite the increasing availability of mobile applications, including global navigation satellite systems. According to a survey carried out by ASFINAG in the autumn of 2013, 95% of truck drivers on the Austrian motorway network do use still the inflexible and old FM radio as one source of information concerning traffic information while they are on the road. The most frequent complaints about an app developed by ASFINAG for Austrian users included the following three points: 1) it was not always in the native language, 2) it was unavailable abroad, 3) it was not as simple as turning on the radio. (for more details on the state of the art in RTTIS see [Váradi et al. 2015]).

3. The solution

In our view, the solution to the problem of multilingual traffic information is to deliver the information to the road users in languages they understand and in the most natural way they prefer, i.e. spoken native language.

3.1 A use case scenario

We illustrate the envisaged solution[1] through the following typical use case scenario. Imagine a Romanian truck driver starting from Bucharest whose destination is Zagreb via Vienna and who speaks only the Romanian language. On crossing the border to Hungary, no longer an administrative border, he is instantly reminded that he crossed a language border for he can only use the Hungarian traffic information service which is currently available only in Hungarian. After the Hungarian part of the journey the driver reaches the Austrian border and is equally helpless since the Austrian traffic information service is also not available in Romanian. The same case is in Slovenia and Croatia. This means that although there are efficient traffic information services operating in the countries he passed through, the driver is not being able to get such vital information as warnings on roadblocks, dangerous weather conditions etc. for the most of his journey. This situation is illustrated in the left picture of Figure 1. The righthand side picture displays the solution we offer for the problem. The same driver would be using a central internet-based multilingual service that would provide the driver with spoken Romanian language traffic throughout the journey. In addition, the traffic news items he would be receiving would be filtered to include only those relevant to his current position and destination.

[1] The detailed project proposal has been developed to implement the technology described in the present paper on a regional scale, called the MORENA project (cf. http://www.morena-project.eu). The backbone of the whole solution is a cloud-based system that provides the machine translation of traffic information. This is called the MORENA service, which allows deployment in a variety of settings. As one possible implementation, the project offers to develop a multi-platform mobile application for mobile devices (smartphones, tablets, etc.), which will be referred to as the MORENA application or MORENA app. The MORENA application is not intended to be the sole deployment of the MORENA service and the MORENA service should be judged in terms of the potential it offers as embedded technology in the ITS domain and should not be evaluated on the merits of the MORENA application alone.

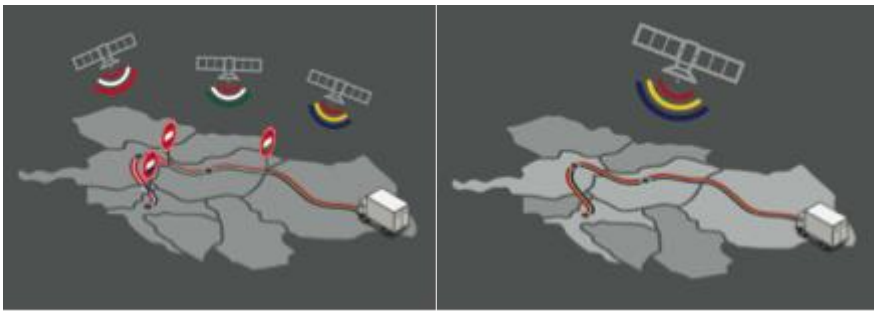

Figure 1 - Typical use case scenarios today (left) and after eliminating language barriers (left)

The technology leverages established technologies in the ITS domain as well as language technologies.

3.2 ITS infrastructure for Real-Time Traffic Information Systems

Information for messages originates from national traffic information centres providing on-line certified information on traffic events. The EU intends to establish national access points for providing co-ordinated and unified traffic information exchangeable among member states.
Messages concerning traffic events are forwarded usually in a strictly coded form. The nowadays traditional way is the TMC in Alert-C coding (EN 2003) that has limited bandwidth therefore somewhat limited data dictionary. Wider bandwidth technologies like DAB or DSRC as well as mobile phone apps use either DATEX II (CEN 2011) or TPEG (TPEG2-TEC 2013) the latter still being in its development phase. DATEX II has a well structured tree-like data dictionary for describing event groups, events and their characteristics. A further strength of DATEX II is the two levels defined for messages: level A has a strict although well extended data dictionary while level B is suitable for tailoring of messages in order to satisfy the needs of different road user groups like transit truck drivers.

DATEX II (CEN/TS 16157-3)			TMC Events (EN ISO 14819-2)			TPEG2-TEC (CEN ISO/TS 21219-15)			
DATEX Class	TYPE	Additional data element	Line	Text (CEN-English)	Code	Cause Code	Sub Cause Code	Warning Level	Text SubCauseCode "trumps" SourceCode
GeneralObstruction	objectOnTheRoad		866	(Q) object(s) on the road. Danger	63	10		3	objects on the road
GeneralObstruction	objectOnTheRoad		863	(Q) obstructions on the road. Danger	902	10		3	objects on the road
GeneralObstruction	shedLoad		868	(Q) shed load(s). Danger	359	10	1	3	shed load
Environmental obstrction	fallenTrees		875	(Q) fallen trees. Danger	906	10	5	3	fallen trees
Environmental obstrction	avalanches		887	avalanches. Danger	992	5	2	3	danger of avalanches
Environmental obstrction	rockfalls		892	rockfalls. Danger	998	9	1	3	rockfalls
Environmental obstrction	landslips		894	landslips. Danger	999	5	4	3	landslips
AnimalsPresenceObstruction	animalsOnTheRoad		944	animals on the road. Danger	923	11		3	animals on roadway
GeneralObstruction	peopleOnRoadway		945	people on roadway. Danger	1482	12		3	people on roadway
GeneralObstruction	childrenOnRoadway		946	children on roadway. Danger	1483	12	1	3	children on roadway
GeneralObstruction	cyclistsOnRoadway		947	cyclists on roadway. Danger	1484	12	2	3	cyclists on roadway
AnimalsPresenceObstruction	largeAnimalsOnTheRoad		948	large animals on roadway	1067	11	4	3	large animals
AnimalsPresenceObstruction	herdOfAnimalsOnTheRoad		949	herds of animals on roadway	1068	11	2	3	herd of animals
DisturbanceActivity	attackOnVehicle		961	people throwing objects onto the road. Danger	897	20	3	4	stone throwing persons
VehicleObstruction	brokenDownVehicle		532	(Q) broken down vehicle(s). Danger	393	13		3	broken down vehicles

Figure 2 - Corresponding Message Subsets for the category "animal/people/obstacles/debris on the road" (TISA 2013)

A well defined part of real time traffic information is the safety related information to be provided free of charge according to EU regulation 886/2013 (EC 2013). In this case the vocabularies of TMC, DATEX and TPEG are more or less similar, as displayed in Figure 2 (TISA 2013).

In case of other traffic events especially incidents DATEX II seems suitable for coding of messages hence there are possibilities to forward the cause of the event as well as its extent in space and time moreover providing the remaining capacity at the road segment or junction concerned.

DATEX II is under continuous development as version 2.3 was released in December 2014 containing three very valuable extensions that support current developments in the user community of DATEX II: mark-up of safety related traffic information situations, parking information and OpenLR as supported location referencing method (information source: http://www.datex2.eu/news/2014/12/01/datex-ii-version-23-available-now).

3.3 The language technology approach

Language technology has now reached a maturity that it can offer a comprehensive full-scale solution to this challenge through a combination of real-time machine translation (MT) and natural sounding speech technology.
We propose here the development of a robust, high quality MT system and its deployment in the ITS domain. We base our confidence on the unique infrastructure of the elaborated terminology and ontology (DATEX II) available for the ITS domain. This language-independent ontology represents a comprehensive and fine-grained conceptual system that allows the description of practically any traffic event and condition. Furthermore, this underlying conceptual scheme is already available translated in different natural languages. The task of machine translation boils down to converting traffic event reports, warnings etc. into a standard DATEX II representation, which can then be mapped into any particular target language and delivered through a text-to-speech system. The technology uses proven components and promises to yield much higher quality translation than is possible through freely accessible general purpose statistical machine translation methods.

The technology described above is amenable to be integrated in ITS solutions in a variety of settings. It provides a generic solution to the multilingual challenge of RTTISs and as such it deserves to be integrated in long-term deployment strategic thinking in ITS. The technology is neutral to modalities of deployment and can be embedded in all sorts of applications providing traffic information. So far we have been focusing on how it can be deployed to automate traditional traffic information services and make them multilingual. In section 3.2. we describe a proposal for a pilot project that aims to implement this technology for five languages and delivers a regional service through an independent online application. This, however, is only one of the possible ways this technology can be employed.
The envisaged technology presents a ground-breaking solution to pre- and on-trip traffic information provision by delivering traffic related information in the most natural and user friendly manner, i.e. in the form of natural, spoken messages in one's native tongue, something like an automated traffic news internet radio. The language technology draws on the standardized data dictionaries and protocols developed in the ITS domain and thus exploits the synergies between the two disciplines involved in ITS.

This innovative approach integrates existing best practices in two domains, the ITS domain and the language technology (LT) domain. Within the LT domain this approach makes use of proven concepts and technology in the fields of terminology in the fields of terminology management, machine translation, computer-assisted translation (CAT) using translation memories (TM), and controlled natural language (CNL) systems. All methods and technologies suggested for this approach have been tested and are robust enough for the real-world applications. What is innovative within the language technology domain in its configuration for this solution is the specific interactive model of three levels/methods/approaches that so far have co-existed with each other, but that have not yet been integrated into a single, coherent and production-oriented framework:

1. terminology management,
2. hybrid process models for MT plus translation memories,
3. CNL authoring approaches.

The specific reasons why this integrative and interactive approach is chosen lie in the inherent limitations of each of these three levels and methods when it comes to overcome cross-lingual barriers in information systems as well as communication processes.

With the new EU regulations on ITS and the already matured DATEX II information interchange format, the context of the LT application is very well specified. The constrained structure of the natural messages, the limited vocabularies (see http://www.datex2.eu/content/datex-ii-v20-data-dictionary), the existing formal multilingual terminologies, and the clear classification of the relevant traffic events allow LT tools to achieve a higher accuracy and much faster processing of the ITS relevant messages than in the case of unrestricted texts. The accuracy and response time are crucial elements in this application domain. We see the DATEX II XML encoding schema as a language independent representation level (interlingua) for the information which could have been initially provided in any of the languages of the project (and whatever other languages in the future). Vice-versa, the messages in DATEX II interlingua representation are source for production of any target language messages (in whatever languages in the future).

As mentioned before, this interlingua-based approach to this translation tasks is facilitated by the limited domain, very well structured and using restricted linguistic constructs and closed vocabularies. It has all the advantages advocated by classics of the interlingua-based MT and resists all early criticism regarding the expressivity and coverage. Moreover, due to the restricted language and precise terminology, one may discard expensive NLP phases (lexical and syntactic disambiguation, discourse phenomena) and use much faster IE techniques (e.g. named entity recognition, regular grammar patterns, event frames). In translating DATEX II messages into natural language, the same specificities of the application domain turn the process into a very fast procedure supported by standardized multilingual patterns and multilingual lexicons and terminologies.

3.4 The data-flow from event reporting to public road traffic news

To illustrate the nature and scale of the MT problem in both the analysis and synthesis phase, we provide a detailed case study of the data flow from the practice of the Hungarian traffic data provider system Útinform.
Real-time events are reported to the relevant county office of Hungarian Public Road Company by phone (police, fire stations, etc.). Details of the event are manually entered through a screen template into a propriety IT system, called KOMVIR (see a sample event in Figure 3). KOMVIR has recently been made compatible with DATEX II standard, thus it is automatically mapped to DATEX II structure (see Figure 4), which is used to exchange data with external partners through the data portal of Útinform.
From the stream of data a subset is chosen for public reporting by a team that is on 24/7 duty at Útinform. Using their competent knowledge of traffic terminology the team converts the structured data to messages ready for broadcasting. They look at the template on the screen, and using additional information (marked with red in Figure 5) they compose a message for the general public.

Road number	From location	Till location	Start point in km	Start point in m	End point in km	End point in m	Date of work start
M0			73	0	76	708	2014.03.26

Time of work start	Rela date of work end	Real time of work end	Scale of limitation	Reason of limitation
11:35	2014.03.26	16:00	2 moving lanes	Works on road surface

Figure 3 - Event recorded in the KOMVIR system (Hungarian original translated into English)

```
<validityTimeSpecification>
      <overallStartTime>2014-03-26T11:35:00.000+01:00</overallStartTime>
      <overallEndTime>2014-03-26T16:00:00.000+01:00</overallEndTime>
</validityTimeSpecification>
<roadNumber>M0</roadNumber>
<linearElementNature>road</linearElementNature>
<startPointOfLinearElement>
      <referentIdentifier>M0 73+0</referentIdentifier>
      <referentName>szelvényszám</referentName>
</startPointOfLinearElement>
<endPointOfLinearElement>
      <referentIdentifier>M0 76+708</referentIdentifier>
      <referentName>szelvényszám</referentName>
</endPointOfLinearElement>
<roadMaintenanceType>maintenanceWork</roadMaintenanceType>
<maintenanceWorksExtension>
      <mkMaintenanceWorks>
            <limitationType>gapFillingWork</limitationType>
      </mkMaintenanceWorks>
</maintenanceWorksExtension>
<supplementaryPositionalDescription>
      <affectedCarriagewayAndLanes>
            <lane>Lane2</lane>
      </affectedCarriagewayAndLanes>
</supplementaryPositionalDescription>
```

Figure 4 - Event report coded in DATEX II (excerpt)

The M0 road leading to the Megyeri bridge general maintenance work will be being done between 73 to 76 kilometers, therefore from 11.30 am to 16:00 pm by the junction of the road leading to Dunakeszi the inner lane will be periodically closed.

Figure 5 - Event in natural language (Hungarian original translated into English)

In Figure 6 we provide a sample of DATEX II data dictionary[2], illustrating the fine level of resolution at which traffic events are captured in standardized format. The DATEX II data dictionary was originally developed in English. As a result of the EasyWay project (http://www.easyway-its.eu/) it has been localized into Hungarian. We expect that such localized data dictionaries will be prepared also in Croatian, where it does not exist yet.

Enumeration name	Enumeration litteral	Designation	Origin	Original code
AccidentTypeEnum	accidentInvolvingRadioactiveMaterial	Accident involving radioactive material	- null -	- null -
AccidentTypeEnum	accidentInvolvingTrain	Accident involving train	- null -	- null -
AccidentTypeEnum	chemicalSpillageAccident	Chemical spillage accident	DATEX	ACE
AccidentTypeEnum	collision	Collision	- null -	- null -
AccidentTypeEnum	collisionWithAnimal	Collision with animal	- null -	- null -
AccidentTypeEnum	collisionWithObstruction	Collision with obstruction	- null -	- null -
AccidentTypeEnum	collisionWithPerson	Collision with person	- null -	- null -
AccidentTypeEnum	earlierAccident	Earlier accident	DATEX	ACA
AccidentTypeEnum	fuelSpillageAccident	Fuel spillage accident	DATEX	ACF
AccidentTypeEnum	headOnCollision	Head on collision	- null -	- null -
AccidentTypeEnum	headOnOrSideCollision	Head on or side collision	- null -	- null -
AccidentTypeEnum	jackknifedArticulatedLorry	Jack-knifed articulated lorry	DATEX	AJA
AccidentTypeEnum	jackknifedCaravan	Jack-knifed caravan	DATEX	AJC
AccidentTypeEnum	jackknifedTrailer	Jack-knifed trailer	DATEX	AJT
AccidentTypeEnum	multipleVehicleCollision	Multiple vehicle collision	- null -	- null -
AccidentTypeEnum	multivehicleAccident	Multivehicle accident	DATEX	ACM
AccidentTypeEnum	oilSpillageAccident	Oil spillage accident	DATEX	AOI
AccidentTypeEnum	other	Other	- null -	- null -
AccidentTypeEnum	overturnedHeavyLorry	Overturned heavy lorry	DATEX	AOL
AccidentTypeEnum	overturnedTrailer	Overturned trailer	- null -	- null -
AccidentTypeEnum	overturnedVehicle	Overturned vehicle	DATEX	AOV
AccidentTypeEnum	rearCollision	Rear collision	- null -	- null -
AccidentTypeEnum	secondaryAccident	Secondary accident	DATEX	ACD
AccidentTypeEnum	seriousAccident	Serious accident	DATEX	ACS
AccidentTypeEnum	sideCollision	Side collision	- null -	- null -
AccidentTypeEnum	vehicleOffRoad	Vehicle off road	- null -	- null -
AccidentTypeEnum	vehicleSpunAround	Vehicle spun around	DATEX	ASP
AlertCDirectionEnum	both	Both	DATEX	B or b
AlertCDirectionEnum	negative	Negative	DATEX	N or n
AlertCDirectionEnum	positive	Positive	DATEX	P or p
AlertCDirectionEnum	unknown	Unknown	DATEX	U or u
AnimalPresenceTypeEnum	animalsOnTheRoad	Animals on the road	DATEX	ANM
AnimalPresenceTypeEnum	herdOfAnimalsOnTheRoad	Herd of animals on the road	DATEX	ANH
AnimalPresenceTypeEnum	largeAnimalsOnTheRoad	Large animals on the road	DATEX	ANL

Figure 6 - A sample from the DATEX II data dictionary

The services in other countries function in the similar way, so we can present the current state of data-flow from event reporting to public road traffic news with a generalised diagram (see Figure 7) and use this diagram to illustrate how MORENA technology will replace and improve the human effort in each step.

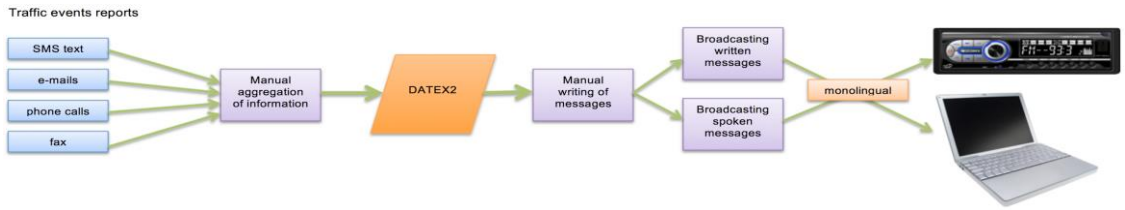

Figure 7 - Generalised diagram representing the current state of event reporting to public road traffic news

[2] http://www.datex2.eu/content/datex-ii-v20-data-dictionary

In Figure 7 it can be noticed that the reports on different traffic events are being collected from different sources (not necessarily digital) and that it includes the engagement of human effort at three different points in the workflow: 1) in aggregation of information, 2) in writing of messages, 3) in speaking out loud the written messages. The existing services for traffic events reports are mostly monolingual and are being broadcasted through RDS channel or through internet either for custom-made client applications or general purpose browsers. In the case of occassional broadcasting of messages in other languages, usually during the summer when the number of foreign drivers is expected to enlarge, the messages are also translated using human effort, but to the fixed and limited number of target languages.
In this and the next diagram (see Figure 8) the boxes in light blue are covering the steps of data acquisition and dissemination, the boxes in violet are the steps that involve heavy human effort and can be replaced by different Language technology modules, while the orange box (DATEX2) is the common internal conceptual representation. Light orange boxes represent either monolingual or multilingual nature of messages that are being broadcasted through different channels.

4. Architecture of MT approach

Deployment of the proposed language technology changes the workflow for traffic events reports (see Figure 10) starting from the sources, i.e. most of them are expected in the form of predefined online templates or free text, but both already in the digital form. However, even the non-digital channels, i.e. spoken messages, can be covered by the Automatic Speech Recognition module that can provide the text from spoken sources (e.g. phone calls) thus eliminating the need for human effort in transcription. The human effort is also not needed for aggregation of information since the incoming free text messages are being preprocessed by standard LT tools followed by automatic multilingual Information Extraction techniques in order to extract the relevant information.

In the next step this information is being converted into the common DATEX II format. The Automatic Text Generation techniques are used for generation of text messages from DATEX II format and then they are submitted to Text-to-Speech module that generates the spoken messages automatically. In these two last steps the human effort is avoided as well, so there is no need for human writers or speakers of messages that are being broadcasted. Once generated in the language chosen by the user and in the spoken form, these messages could be broadcast either through the existing navigation devices (providing that they are compatible with the proposed MORENA technology), or through the special client in the form of the MORENA app that runs on the most popular mobile operating platforms (Android, iOS, Winphone).

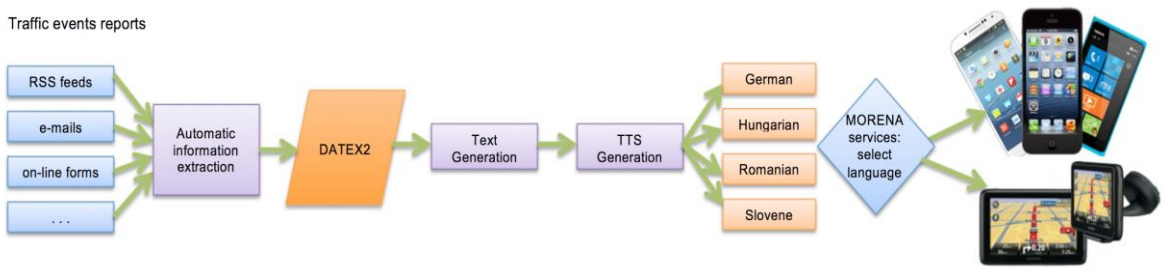

Figure 8 - The data processing workflow using MORENA technology

The architecture of the proposed system regarding the data processing workflow can be viewed as a chain of three logical modules (see Figure 8): data acquisition (input), data storage, and data distribution (output). Language technology is present in both the input and the output modules. The input module handles data acquisition and transformation into the standard DATEX II structured format. Data sources are both semi-structured (non-DATEX XML/RSS feeds) as well as unstructured (in the shape of free natural language text, e.g. in articles regarding traffic news, or even the phone calls). The output module handles the translation from the semi-structured format into natural language texts that users will receive through the MORENA app or any other compatible way in their native languages.

Language technologies are applied to a number of essential tasks:

1. *Transforming from a non-DATEX XML format to a DATEX format (input).* A number of traffic information websites offer XML formatted feeds (ex: RSS) regarding traffic events. In most cases, a simple mapping between the XML feed and a DATEX event is sufficient. More complex situations do require the usage of multilingual terminological lexicons together with standard regular expression extractors and rule-based mappings to completely match a certain XML traffic event to the DATEX format.

2. *Extracting DATEX formatted traffic events from natural language articles (input).* Real-time traffic information is also provided in the form of natural language text in traffic announcement articles. A number of advanced language technologies take the shape of an Information Extraction module that parses free text and produces the semi-structured DATEX traffic event. The text will first be pre-processed: split into sentences, tokenized, lemmatized, POS tagged. Tools that provide these services with high accuracy exist for each of the covered languages, e.g. (Halácsy et al. 2007) and (Tufiş et al. 2008). Pattern and named entity recognizers will mark important anchor points like cities, highways, junctions, date/time, values, as e.g. (Agić–Bekavac 2013) and (Agić et al. 2010). We will use a custom frame-based IE method tailored to traffic events. Frames will represent traffic event categories, each having their unique attributes such as how many lanes are affected in a road repair event or meteorological data for a road closure event. Having the annotated free text (POS tagged, entity marked, shallow relations between entities extracted, etc.) appropriate frames will be filled with relevant info. At this point, a simple mapping between a frame and its DATEX representation will be all that is required to successfully complete the IE step.

3. Translating DATEX formatted traffic events to natural language text in each of the project's selected languages (output). Controlled natural languages (CNLs), also called "simplified" or "technical" languages, are subsets of natural languages, obtained by restricting the grammar and vocabulary to reduce or eliminate ambiguity and complexity, and have gained increasing importance and relevance for a varied number of applications (e.g. [Kuhn 2013] and [Kuhn–Fuchs 2012]).

In this specific context this approach and thus this component is relevant and useful because of the highly regular/ritualized nature of traffic messages. DATEX and related ITS standards/specifications include already standardized typologies of traffic messages (road accidents, weather conditions, traffic jams, etc.) and due to corpus analysis of existing recorded traffic messages such typologies can be mapped to discourse patterns in each

language concerned that are actually used in such traffic messages. If the experience from automatic text-to-ontology mapping attempts (e.g. [Buitelaar–Cimiano, 2008]) will be proven to be helpful in this narrow domain, it might be useful to convert the existing DATEX dictionary into an ontology format (e.g. OWL or similar). This will open a whole new perspective of DATEX reusability in the knowledge processing systems. In addition, there are preferences by TTS modules to standardize the structure of traffic messages in the form of controlled language which even further facilitates the suggested approach described here.

The translation will follow a standardized procedure: entities from DATEX structures will be extracted, and, based on the event type, will be placed as anchor points in future sentences. Depending on the relations between them, pre-defined text will be generated, filling in the missing spaces between the entities.

The generation of natural language text in all the different languages will require the use of CNL corpora, multilingual terminological lexicons and a translation system for each target language pair. Language models can be used to ensure grammatical correctness in the form of number, gender, case and person agreements, verb tenses, etc. (for the envisaged languages these grammatical features are highly relevant). The paragraph structure will also follow a predefined discourse pattern, making the overall generated natural language text concise, fluent and easily understandable, ready to be sent to the TTS synthesizer.

Speaking about MT in closed domains brings to the discussion one well-known dichotomy in MT (especially of phrased-based statistical approaches): in-domain vs. out-of domain translation and several studies showed large quality differences in favour of the in-domain translations (Tufiș et al. 2013). The dichotomy in-domain/out-of-domain refers to the types of the texts to be translated: if they are similar (lexically, syntactically and thematically) to the data used for training of the translation and language models, translation is called in-domain. Models adaptation/extension is also significantly more productive when the in-domain translation assumption remains valid as demonstrated by recent experiments (Dumitrescu et al. 2013). The same considerations hold for text-to-speech synthesis. The MT translation and TTS as envisaged here are typical in-domain settings translation and TTS processes.

The MORENA technology as described here addresses the limitations noted above in RTTIS practice on several counts:

1. It will eliminate language barriers (see Figure 1) through high quality MT of traffic related information from and into the national languages of the participating countries (plus English offered as a lingua franca to speakers of other languages until the service is up-scaled to include their language).
2. It will deliver the traffic information in spoken language, the most natural medium. This, as opposed to reading a written message in whatever device, is an obvious requirement from the perspectives of safe driving.
3. It will cover traffic information that is of direct relevance to drivers both in terms of their geolocation point and their destination. This personalized and user oriented traffic information service has a clear advantage over current practices and on top of that it is delivered in the medium, spoken news, that drivers predominantly prefer while driving.

4. This user customized service is delivered through the Internet to their personal mobile devices, typically smartphones, which are becoming widespread on a massive scale. The viability of this service is ensured with the planned abolishment of roaming charges within the European Union.

5. Deployment

In addition to the innovative core multilingual language technology employed, the anticipated deployment of the suggested service will be innovative as well. It will be a cloud based service delivered as a mobile application for smart phones or other mobile devices. The recently announced phasing out of roaming charges from 2015 represents a major breakthrough leading to mobile devices as the main broadcasting channel for traffic information services (TIS). An online service has the additional benefit over current practice that the information flow through this channel can be tailored for individual users since all relevant information such as preferred language, current GPS position, planned destination, etc. are at disposal. So delivery of information is personalised not only for language, but also for GPS position and destination.

5.1 System architecture

The MORENA project will be developed using a clear client-server architecture, with mobile (smartphone) app acting as client, backed by a scalable cluster of interconnected services in the server back-end.
Mobile apps are to be developed in three currently most popular mobile platforms – Apple iOS platform for iPhone and iPad devices, Google Android platform available on a vast array of smartphones and tablets of different manufacturers, and Microsoft Windows Phone platform as a third competitor in the smartphone and tablet market.
The app on all three platforms would be targeted specifically for smartphones, as it is our intended use case scenario, but should also work on tablet devices in compatibility mode (since they would not utilise phone-only features, such as placing calls).
Since there is no supported way of programming a single, platform independent mobile app codebase and distributing it to various mobile platforms, a total of three separate apps with the exact same functionality are to be developed, one for each mobile platform. Therefore, regardless of the platform, all apps should provide the same set of features, differing only in matters of user interface design and interaction styles used, which is to be tailor-made to the specific design and interaction guidelines of each mobile platform. In turn, that would ensure highest-quality user experience and customer satisfaction.
Since MORENA will use different TTS engines for different languages, and as these systems are proprietary it is conceived unreasonable to re-develop these systems to fit on a mobile device, especially given that multiple, generally incompatible mobile platforms are to be supported. Therefore, TTS engines would need to reside on a server, and in order to provide all the proposed features, the app would rely heavily on mobile data usage, as well as location services (GPS).
It has been taken into account that by December 2015, the mobile roaming charges would be dropped in the EU, rendering mobile data-intense apps viable for common use even across national borders within the European Union.

Because of the ever-changing nature of traffic and travel information, the app would not need to store any data locally on the device, with the exception of basic configuration settings.
In order to download only the data relevant to travel, user's location is sent to the server, so a spatial query against relevant and available traffic and travel messages from the database could be performed.
All the travel and traffic data displayed in text, as well as audio sequences reproduced would first need to be downloaded or streamed in real-time from the Internet, more specifically from a MORENA web service endpoint, an edge node in the MORENA server back-end.
Due to the architectural and bandwidth constraints of the mobile platform, a simple RESTful web service based on JSON messages is to be used for data exchange.

As the client app would only need to display the relevant data and reproduce the streamed synthesised audio sequences, the general MORENA framework and majority of services need to reside on a server back-end (see Figure 9).
Although the complete server back-end for a proof of concept scenario could be implemented on a single web server, it would not be recommended as significant load could be generated easily by multiple users, given the complexity of the solution and some CPU-intense operations, such as synthesising voice. Therefore, a distributed system is introduced consisting of various server roles, where each can be implemented on multiple server nodes for scalability:

- database role (DB server): stores a complete or partial copy of travel and traffic messages received by service providers;
- text-to-speech synthesiser role (TTS server): generates audio from natural language input to be streamed to devices; for large-scale deployment multiple instances would need to be provisioned;
- client access role (CAS server): acts as a web service endpoint covering multiple data exchange tasks:
    - receives the DATEX II messages from service providers and stores them into appropriate DB servers;
    - generates personalised traffic info reports based on user's location, speed and heading, in both natural language and as structured data, from the stored DATEX II messages, to be delivered to the end user devices;
    - initiates TTS conversions based on latter and delivers the audio stream to the user.

In order for the system to receive messages, service providers must first make their traffic information management solutions compatible with the DATEX II format and implement a continuous stream of messages towards the MORENA endpoint (CAS).

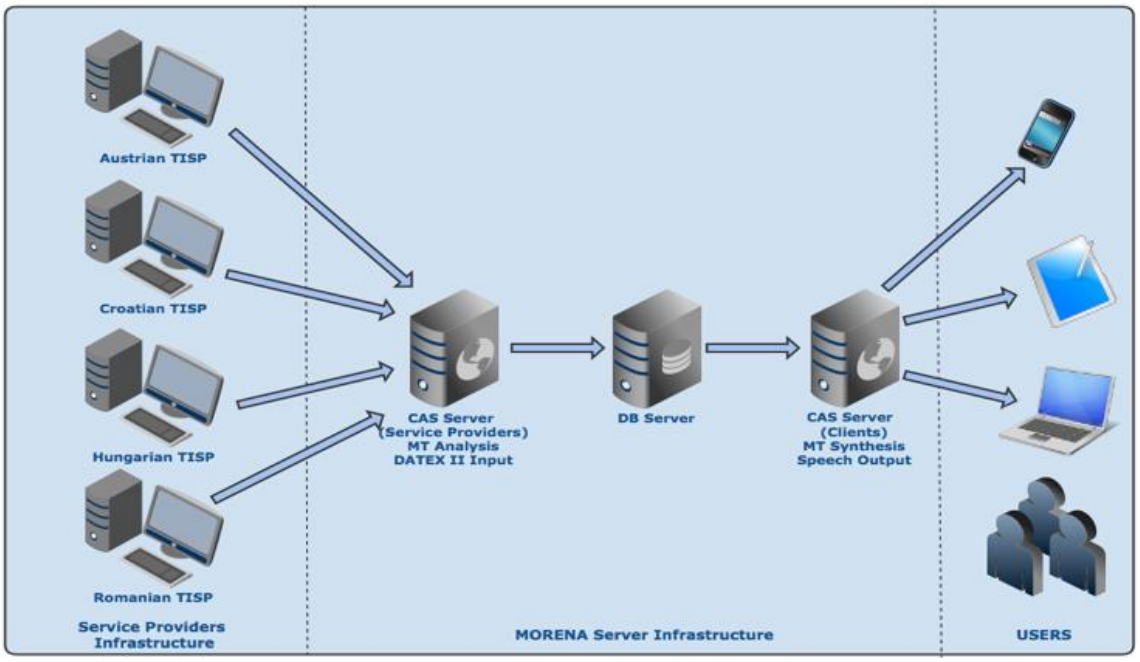

Figure 9 - MORENA Server Infrastructure in relation to users and service providers

5.2 MORENA vs. GNSS applications

It is not the task of this paper to justify current traffic information services or discuss how they relate to GNSS based services. We merely note that the information that traditional traffic information services provide partly overlaps, partly complements what one finds in navigation systems and there is no reason why navigation systems should not integrate the information services provided by the traditional service providers. Appendix 1 contains a comparison of RTTIS and GNSS systems in terms of spoken data. Appendix 2 displays a detailed comparison of the kind of content the two systems typically transmit. It is easy to see how the two systems usefully complement each other. Navigation systems are great in indicating up-to-the-minute traffic conditions. However, drivers may well be interested in the reasons why the particular congestion ahead has developed, how long it is estimated to last, etc. An important part of traffic information consists of warnings about road conditions, weather, animals, etc. that navigation systems currently usually do not report. As was discussed earlier, current GNSS based applications operate in various languages but the localisation is limited to the user interface and the limited set of predefined and usually pre-recorded navigation messages. If information about real-time traffic is conveyed at all, it is presented mostly visually with all the evident limitations. In conclusion, we can state that to the extent the whole spectrum of traffic information is integrated in GNSS based applications, possibly as an optional component, the spoken automatic Machine Translation of messages can be successfully deployed in the navigation systems. In this respect the proposed technology can be seen not as a rival to GPS navigation systems, but as complementary technology that can be included in existing GPS navigation systems and make them more useful and adaptable to their user's needs.

Another area to consider is built-in car traffic information systems. Here again we foresee no major problems in deploying this technology. Although it is inherently based on DATEX II as the conceptual backbone, its output is digital speech, which can be deployed in TPEG-based information systems as well.

Conclusions

The technology offers an additional component to the existing suggested solutions, e.g. it can be embedded or connected into existing GPS navigation systems.
It is oriented towards digital audio broadcast (e.g. similar to user-tailored internet radio) as dominant channel for traffic information delivery.
Some messages are important warnings to drivers (road conditions, weather conditions, traffic jams, accidents, etc.) and they should be expressed in the most natural way.
Drivers prefer to ***hear*** about the road conditions while driving and ***audio*** is their ***preferred channel***.
The suggested multilingual technology is neutral and applicable to existing recommendations and current ITS solutions. The objectives of the described service perfectly align with the strategic aims of the ITS Directive (EC 2010) The workflow is entirely compatible with the value chain defined for the Traffic Information Service in a position paper of the Traveller Information Services Association (TISA 2014).

References

1. Agić, Ž., Bekavac, B. (2013). Domain-aware Evaluation of Named Entity Recognition Systems for Croatian. *Journal of computing and information technology,* 21 (3), pp. 195–209.
2. Agić, Ž., Šojat, K., Tadić, M. (2010). An Experiment in Verb Valency Frame Extraction from Croatian Dependency Treebank. In *Proceedings of the 32nd International Conference on Information Technology Interfaces,* Cavtat, Hrvatska: University of Zagreb. pp. 55–60.
3. Buitelaar, P., Cimiano, P. (2008). *Ontology Learning and Population: Bridging the Gap between Text and Knowledge.* Amsterdam: IOS Press.
4. CEN 2011: CEN/TS 16157-3:2011. *Intelligent transport systems - DATEX II data exchange specifications for traffic management and information,* Situation publication.
5. Dumitrescu, Ș. Ion, R. Ștefănescu, D. Boroș, T. Tufiș D. (2013). Experiments on Language and Translation Models Adaptation for Statistical Machine Translation. In Tufiș, D. Rus, V. Forăscu, C. (eds.) *Towards Multilingual Europe 2020: A Romanian Perspective,* pp. 205–224.
6. EC 2010: European Commission (2010). *Directive 2010/40/EU of the European Parliament and of the Council*. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1438350993825&uri=CELEX:32010L0040 (last visited 01-08-2015)
7. EC 2013: *Commission Delegated Regulation (EU) No 886/2013 of 15 May 2013.* Available at http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32013R0886 (last visited 01-08-2015)

8. EN 2003: EN ISO 14819-2:2003. *Traffic and Traveller Information (TTI). TTI messages via traffic message coding Event and information codes for Radio Data System. Traffic Message Channel (RDS-TMC).*

9. Halácsy, P. Kornai, A. Oravecz, Cs. (2007). HunPos - an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic: Association for Computational Linguistics. pp. 209–212.

10. Kuhn, T. (2013). A Principled Approach to Grammars for Controlled Natural Languages and Predictive Editors. *Journal of Logic, Language and Information,* 22 (1).

11. Kuhn, T. Fuchs, N. E. (2012). *Controlled Natural Language.* Springer.

12. TISA 2013: Traveller Information Services Association (2013). *Safety related message sets – Selection of DATEX II Codes, TPEG2-TEC-Causes and TMC-Events for EC high level Categories.* Available at http://www.tisa.org/assets/Uploads/Public/ITSTF13004SafetyrelatedMessage-Sets-DATEXII-TPEG-TECandTMCv3.pdf (last visited 03-08-2015)

13. TISA 2014: Traveller Information Services Association (2014). *TISA Position concerning a public consultation of the European Commission on the Provision of EU-wide Real-time Traffic Information Services.* Available at http://www.tisa.org/assets/Uploads/Public/EO14005TISAPosition-paperPriority-Action-Bfinal_4.pdf (last visited 02-08-2015)

14. TPEG2-TEC 2013: *Intelligent Transport Systems (ITS) — Traffic and Travel Information (TTI) via Transport Protocol Experts Group, Generation 2 (TPEG2) - Part 15: Traffic Event Compact.*

15. Tufiş, D. Ion, R. Ceauşu, A. Ştefănescu, D. (2008). RACAI's Linguistic Web Services. In *Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008,* Marrakech, Morocco: ELRA.

16. Tufiș, D., Ion, R., Dumitrescu, S. D. (2013). Wiki-Translator: Multilingual Experiments for In-Domain Translations. In *Computer Science Journal of Moldova,* Academy of Sciences of Moldova, Institute of Mathematics and Computer Science, ISSN 1561-4042, vol.21, no. 3(63), 2013, pp. 332–359.

17. Váradi, T. Tadić, M. Gulyás, A. Niculescu, M. (2015). *Language Technology in the Service of Intelligent Transport Systems*, Paper number ITS-2878, 22nd ITS World Congress, Bordeaux, France, 5–9 October 2015.

Appendix 1

Navigation devices and spoken RTTI

Name:	Travel Pilot
Manufacturer:	Blaupunkt
Maps:	TomTom
TMC:	Yes, information displayed on screen as list of events and considered for dynamic routing.
TTS:	Only for street names and guidance.

Name:	HERE Maps
Availability:	Available as HERE Maps and HERE Drive apps for Windows Phone, Android and iOS smartphones.
Display:	Real time traffic information displayed colour coded and considered for dynamic routing.
TTS:	Only for street names and guidance.

Name:	Nüvi
Manufacturer:	Garmin - https://buy.garmin.com/en-US/US/on-the-road/automotive/2013-line/nuvi-2598lmthd/prod122523.html
Maps:	NAVTEQ/HERE
Traffic:	TMC-FM/DAB and dedicated services over the Internet via smartphone (Android and iOS) Bluetooth link.
Display:	Displayed on screen as list of events, colour coded, spoken and considered for dynamic routing.
TTS:	Traffic including events, street names and guidance.

Name:	iGo
Application:	iGo Primo app for iPhone - http://www.igonavigation.com/igo-my-way-for-iphone and also built-in other devices - https://en.wikipedia.org/wiki/IGO_(software).
Display:	Real time traffic information displayed colour coded, as alerts, spoken and considered for dynamic routing.
TTS:	Traffic including events, street names and guidance.

Various smartphone apps	
Example 1:	https://play.google.com/store/apps/details?id=net.monthorin.rttraffic
Example 2:	https://play.google.com/store/apps/details?id=com.glob.plugins.tts.
Service:	Possibly it only gives traffic status.

Appendix 2

Table 1 - Comparision of GNSS systems with RTTI services

Category of traffic information	Navigation devices	RTTI (web)services
Traffic speed flow	Available, if equipped with RDS/TMC receiver or DAB/TPEG receiver or Internet connection. Only current situation can be displayed.	Available. Historical and predicted information can also be provided.
Traffic events	Available, if equipped with RDS/TMC receiver or DAB/TPEG receiver or Internet connection. Typically displayed as a pictogram or as a list, with minimal details.	Available. Detailed information is provided, for example: source of the data, duration of the event, description, timestamp of last update.
Road weather events	Available, if equipped with DAB/TPEG receiver or Internet connection.	Available.
Weather forecasts	Available, if equipped with DAB/TPEG receiver or Internet connection.	Available.
Animal warnings	Available, if equipped with DAB/TPEG receiver or Internet connection.	Available.
Special events	Available, if equipped with DAB/TPEG receiver or Internet connection.	Available.
Police check points	Available, no data connection required.	Availability depending on the baseline map used.
Road tolls	Available, no data connection required.	Availability depending on the baseline map used.
Parking places	Available, no data connection required.	Availability depending on the baseline map used.
Webcams	Not available. Just position of speed cameras might be provided.	Available with live video feed.
Bike stations	Available, no data connection required.	Availability depending on the baseline map used.