1. Publishable summary

1.1 Summary of Project Description and Objectives

In the last two decades, interactive spoken dialogue systems (SDS) have moved from research prototypes to real-life commercial applications. SDSs are now pervasive and represent a business with yearly revenue of over 1B Euros worldwide. Still, one major roadblock in commercial SDS prototyping is that they are not easily and quickly *portable to new domains or languages.* Currently, porting spoken dialogue systems to a new application domain demands significant effort and expertise for: defining a domain ontology, handcrafting grammars for recognition and understanding, designing the dialogue state-machine, designing and recording of prompts. For example, for a medium complexity speech service, about 20-25% of the prototyping effort goes into design of static and dynamic grammars. In order to extend the system to a new language, again major expert effort is needed. General-purpose language resources that are built top-down by linguists often prove inadequate for SDS design. A major shortcoming of existing linguistic resources is the lack of coverage, e.g., proper nouns are not adequately covered in the European WordNets despite their high rate of occurrence in the web. Furthermore, existing resources do not provide appropriate text data to accompany lexical entries; for SDS tasks, e.g., training of statistical grammars, *data is equally important to semantics.*

In PortDial we bring together European SMEs that are developing state-of-the-art spoken dialogue systems and the handcrafted semantic components underlying such systems with research institutions at the forefront of progress in the automatic creation or enrichment of semantic language resources. Our aim is to *apply these technologies towards the creation of domain-specific multilingual SDS resources*, specifically, *data-linked ontologies and grammars.* The project will result in 1) a commercial platform for quick prototyping of interactive spoken dialogue applications in new domains and languages, and 2) a multilingual collection of corpus-linked ontologies and grammars for these domains and languages (both commercial and free version).

PortDial is built around the *knowledge cascade of technologies, data and services.* Automatic or machine-aided algorithms will be used to create linguistic resources for SDS, and, in turn, these data will be used to create cost-effective speech services and platforms. Data is at the centre of this cascade, thus, linguistic resources are the centrepiece of PortDial. The main S&T goal of PortDial is to *devise machine-aided methods for creating, cleaning-up and publishing multilingual domain ontologies and grammars for spoken dialogue system prototyping in various application domains.* We have selected a **machine-aided** (human and machine in the loop) approach because it is less disruptive (wrt to legacy processes) and may lead to optimal closed-loop process performance. We want to create resources not only for popular domains and languages, where linguistic resources (e.g., ontologies) might be available, but *also for relatively under-resourced domains* (e.g., concert ticketing) *and languages,* (e.g., Greek). As discussed next, a different approach is taken algorithmically for each the resource-rich and resource-poor scenarios.

PortDial adopts a **user-centric approach** to SDS resource building. Rather that simply rolling out resources from the research lab to the real world - being hopeful about their usefulness, - we have tried to map (see also previous section) the requirements of a speech services developer and emulate the logical flow being followed. In doing so, we have identified two scenarios for resource building: **domain porting** where the developer starts from an existing application in one domain and ports the resource to a new (related) domain, and **language porting** where the application is ported to a new language. The main difference between the two scenarios is that for domain porting the emphasis is on **adapting** the language resources, while for language porting the emphasis is on **translation.** Thus our second goal is to *create a platform that supports cost-effective language resource building for those two scenarios: domain and language porting, as well as, use this platform to prototype and evaluate speech services.* The platform will also include interfaces for iterative correction and post-editing of the linguistic resources.

Although providing linguistic resources and a platform for cost-effective SDS development is important and relevant, a data pool that is not being updated and enriched quickly fails its purpose. It is thus important to guarantee the sustainability of the linguistic SDS resources engineered in PortDial; to do so a community of users has to be put together that (does not only use but also) develops further the data pool. Towards this end,

we have taken a multi-pronged exploitation approach by supporting both a free and premium data exchange, as well as, commercializing the speech services platform. Our third goal is to *create and support a sustainable pool of users that contribute to a linguistic resources data exchange.* Two separate groups of users are targeted: non-commercial users including the research community that can maintain and enrich the free version of the data pool, and commercial speech services developers that can contribute to the premium data pool through crowd-sourcing (bartering) or in an electronic marketplace. Towards this goal, we intend to **exploit existing communities** for sharing linguistic resources, such as, META-NET, and the corresponding data sharing infrastructure META-SHARE, http://www.meta-net.eu/meta-share. It is expected that the open data pool can also stimulate research interest in automated linguistic resource creation, also via targeted actions of PortDial, such as, the evaluation workshop and the user conference.

1.2 Achievements and Main Results

Next we outline the achievement and main results in Y1 of the PortDial project as it pertains to the three main technical goals: ontologies, grammars, and localization (porting of resources to new languages), as well as the main technological, namely the creation of a machine-aided solution for the crafting of linguistic resources for new SDS domains and language (i.e., the PortDial platform).

First we present the results of our work towards the creation of the **ontology** (conceptual space) of Spoken Dialog Systems (SDS) from a corpus, i.e., the semi-automatic induction of domain specific ontologies (or taxonomies) from a corpus, as well as the population of these ontologies with specific instances. In addition, we present progress on a wide-range of core technologies used for grammar induction and localization, specifically: 1) semantic similarity estimation, 2) named-entity detection, 3) web-data harvesting and annotation, 4) lexicalization of ontologies for top-down grammar induction. All these tasks are building blocks for grammar induction and can be used for speeding up SDS development and the efficiency of the resulting SDS systems.

1. Automatic induction of ontologies from text is based on TSI's Ontogain system, which was adapted for SDS data and enhanced with a GUI component enabling user feedback. Ontogain is used for providing bootstrap domain ontologies. Evaluation on air travel and tourism domains indicated that Ontogain's overall performance (F-measure) is in the range of 0.3-0.5. Ontology population (i.e., extract instances of ontology concepts from corpora) was performed by analyzing domain specific corpora using custom rules for detection of references to domain concepts and properties. Several datasets for the air travel and tourism/entertainment domains, existing and web harvested, were analyzed and the accuracy of analysis was evaluated. F-measure in the range of 0.7-0.9, was achieved. When developing a grammar, an existing populating ontology can be a useful resource for suggesting additional concepts and instances. This was achieved by mapping existing grammar concepts to the domain ontology, thus extracting related ontology concepts for addition into the grammar. Promising preliminary results were obtained and the algorithm was integrated in the PortDial platform.
2. Semantic similarity computation is an important building block for grammar and ontology induction. Several metrics of semantic similarity are defined in this report, organized into two broad categories: (i) word-level, and (ii) phrase-level metrics. The word-level metrics were evaluated for the task of noun similarity rating, for which the correlation with human scores was used as evaluation metric. The evaluation was conducted for English, Greek, and German. High correlations scores (> 0.85) were achieved for the majority of English datasets, while lower performance (0.65) was observed for the other languages. The phrase-level similarity metrics were evaluated for two tasks in English, namely, paraphrase recognition, and sentence similarity rating using well-established datasets. High F-measure (0.80) was achieved for paraphrase recognition, while moderate correlation (0.64) was obtained for sentence similarity rating.
3. The creation of domain specific datasets is an important part of the WP2, thus semi-automatic, automatic and crowd-sourcing based approaches are proposed and evaluated for the web harvesting of data for various domains (travel, tourism/entertainment) and languages (English, Italian, Greek). The web- harvested data are evaluated using the domain grammars in terms of richness and

relevance (in- domainess) with very good results. The web-harvested corpora were also used/evaluated for grammar induction/localization (see WP3/4).

4. Finally, automatic grammar induction for SDS based on lexicalization of existing ontologies (corresponding to a resource rich scenario) is presented and evaluated. Evaluation results indicate considerable increase in performance when the user feedback is incorporated.

Towards our goal of automating or machine-aiding the process of generating **grammars** for spoken dialog systems we have performed the following tasks:

1. We have implemented and evaluated a knowledge-based, top-down approach for automatic grammar induction that can be applied in a resource rich scenario
2. We have implemented and evaluated a corpus-based, bottom-up approach for automatic grammar induction that can be applied in a resource poor scenario.
3. In addition, we have developed in Y1 strategies for combining grammars that were generated by the two approaches (top-down and bottom-up), in order to arrive at a fused grammar that outperforms the single grammars, ideally combining the strength of both approaches while avoiding their weaknesses.
4. We have also built a module for evaluating grammars (both top-down and bottom-up grammars), as well as the resulting fused grammars. Results of a currently undergoing evaluation serve to understand the strengths and weaknesses of both approaches, as well as a first strategy for merging grammars. As expected, bottom-up grammars reach a higher recall than top-down grammars, while the latter achieve a better precision. As a first fusion strategy we implemented a simple union of the grammars resulting from the top-down and bottom-up grammar induction. First experiments show that already this simple strategy of late fusion leads to a grammar that improves on the single input grammars. The plan for the second year is to also explore possibilities of earlier fusion strategies.
5. In accordance with PortDial's general objective to identify natural and effective points for human intervention in the automatic generation of SDS resources, we have devised tools for the manual validation and enhancement of automatically induced grammars. This "human-in-the-loop" paradigm of incremental grammar enrichment serves as the basis for the grammar induction algorithms integrated in the PortDial platform in Y1.
6. We have also initiated an effort for SDS prompt enrichment using paraphrasing technology.

For **porting SDS resources across languages** we used machine-aided translation with crowd-sourced post-editing of the resources (annotated corpora) needed to train language-understanding models and build grammars using bottom-up methods. In addition we applied top-down, bottom-up and fused grammar induction approaches on the translated corpora and evaluated the resulting grammars. Finally, these algorithms are to be integrated into an interface for porting dialogue system resources across languages. Specifically:

1. SDS language porting, on a grand scale, is split into Test-on-Source and Test-on-Target scenarios, with respect to the direction and the object of translation. In the former, user utterances in the target language are translated to the language of the existing system; and its SLU is “extended” using statistical machine translation (SMT) to cover a new language, and the success relies on the high quality *machine translation.* In the latter, the data used to build a source SLU or induce grammars is translated to a target language, and new understanding components are created. The second approach relies on the accurate *transfer of annotation* from the source to the target language, in the case of stochastic SLU models, as well as accurate *machine translation.*
2. Test-on-Source scenario was extensively experimented, comparing three SMT approaches: in-domain and out-of-domain data trained moses-based systems, and general- domain off-the-shelf SMT. Pre- and post-processing techniques for both off-the-shelf and out-of-domain SMT were developed and evaluated. The in-domain data trained SMT system produces the best SLU performance, not very far from the performance of the original SLU tested on source language (Concept Error Rate of 25.6 for close-to-source resource-rich language - Spanish, and of 21.5 for the source language - Italian). However, the results for the Turkish, that represents distant and resource-poor language, are not as good. Domain adaptation of out-of-domain moses-based SMT systems is

not very far below the in-domain system in terms of BLEU score; and domain adaptation will continue in year two of the project.

3. For linguistic resource porting, crowd-sourced translation methods were investigated. Methods for quality control of workers as well as for their motivation on NLP task were developed. Additionally, experiments combining human and computer processing were conducted, such as ROVER over several worker judgments. It was observed that for resource-poor languages, it is hard to exploit crowdsourcing platforms for direct translation, as their worker base lacks bilingual speakers for both languages of interest. Thus, the attention was shifted to monolingual tasks for resource-rich languages (Spanish), such as translation ranking. In order to experiment with crowd-sourced tasks for resource-poor languages, targeted crowdsourcing is being explored.

The main focus of our **integration** work in Y1 was to: 1) specify the user requirements and overall architecture of the linguistic resource modules in the spoken dialogue development platform, 2) integrate the ontology evolution and grammar induction modules into the platform, 3) evaluate the initial version of the platform. More specifically the main achievements of Y1 are:

1. The Design & Implementation of the Baseline PortDial Platform (PDP). It provides a web-based interface to basic grammar development, which includes a grammar editor, test case manager, simple grammar evaluation subsystem, grammar visualization capability, and a basic versioning system. A PDP prototype is available on the cloud for testing and evaluation
2. BEM1, a prototype of the bottom up terminal enhancement methodology. Code named BEM1, it enables grammar developers to induce terminal concepts by example: i.e. they enter an example of a city concept, and let BEM1 to propose other similar concepts, taken out of a corpus suitable for their particular domain. The BEM1 is currently locked in to the Travel Domain, but other corpora have been also tested.
3. The Enhancer add-on to PDP (using a bottom up semi-automated approach). Our platform has been extended to incorporate a suggestion interface suitable for both terminal and non-terminal enhancement. PDP integrates the TUC defined bottom up enhancement process with its basic grammar development platform. It also provides some hooks to the ontology subsystem.
4. Although not integrated in yet, we have sketched the workflow for the language porting subsystem of PDP, and we are in the process of its design and prototyping.
5. A detailed definition of the evaluation process in terms of data, procedures, tasks, and metrics. A preliminary evaluation of the PDP shows encouraging results.

The objectives of the **dissemination/exploitation** WP6 are to: 1) achieve widespread awareness about PortDial to all relevant parties (industry, academia, user communities, other EU projects), 2) advertise and promote PortDial scientific and technological achievements at trade-shows, conferences and other events, 3) exploit the PortDial data pool and service creation platform via user communities, B2B and B2C business models, and 4) manage the PortDial intellectual property to maximize exploitation opportunities of PortDial outputs. The main achievement of Y1 are: 1) the creation and population of the project website, 2) the creation of the grammar e-shop and the data sharing portal, 3) the participation of the research and industrial partners in numerous dissemination events advertising the PortDial project and given demos/presentation, 4) the drafting and first execution steps of the exploitation strategies especially for the SME partners and 5) the management of the intellectual property (especially annotated data and grammars) created by the consortium. The main concept behind exploitation strategies of the SMEs which is the central objective of this WP are outlined next.

Overall, as we move to the second year of the project, this report documents our main results and sketches the road ahead. We have built a first version of the platform to provide basic grammar development capabilities, an initial grammar enhancement methodology and prototype, and an evaluation protocol to measure the effectiveness of our strategies. In the next year, we are planning to improve on our grammar enhancement process and to introduce a language porting capability into the system towards.

1.3 Expected Final Results and Impact

Based on the project plan and the aforementioned progress in Y1, the PortDial project is expected to achieve all three objectives by project end. The main outputs of the PortDial project are: 1) A multilingual speech service authoring platform for grammar and ontology authoring for SDS. 2) The concepts-services-grammars multilingual, multi-domain data will be made commercially available as a separate package. The target group here is developers that wish to use the data for prototyping speech services but not necessarily using the PortDial platform. VoiceWeb will make the PortDial speech services platform and associated linguistic resources commercially available. 3) The multilingual domain ontologies, lexica, and associated text data mined from the web will be made available via a CreativeCommons (CC-BY) license allowing their use for non-commercial purposes. These data is mainly targeted to the research community for further developing algorithms for the automatic creation of grammars and, in general, resources for SDS. 4) Open-source software packages for linguistic resources creation provided by the research partners.

PortDial addresses an important business opportunity in the area of speech services and associated linguistic resources. Language resources are the main bottleneck for the quick prototyping and porting of speech services across domains and languages. The PortDial linguistic resources will act as an enabler, filling the void in SDS-specific linguistic resources, lowering the barrier to entry for Europe's SMEs, as well as, improving the quality and cost-effectiveness of speech service prototyping. To maximize impact, we have selected a three-pronged exploitation approach in PortDial for packaging and marketing language resources for SDS: 1) a common license (free) linguistic resources package for research and academia, 2) a commercial linguistic resources package for SDS development in specific domains/languages and 3) a speech services prototyping platform that contains the linguistic resources, as well as, tools for quickly creating such grammars for new application domains and languages. This way the needs of three different user-communities for these resources and technologies are being addressed. The PortDial platform and premium data will be positioned as a product enabling rapid and cost-effective porting of voice applications into new application domains and languages. The target segment for this product will be SMEs worldwide in the mobile application development industry lacking the expertise/resources to develop multilingual speech services in-house. By creating a community of users for the premium data that will contribute resources for new languages and post-edit existing resources, the sustainability of the data pool beyond project end is achieved. The tandem offering of platform/data allows for synergistic exploitation, further enhancing data pool sustainability. Last but not least the free data will be managed by an active user community that will include researchers and speech services developers, following an open-source model, bringing about innovation and creating new market opportunities.

In addition to the impacts listed above, PortDial is also expected to have impact towards the research community by 1) further demonstrating the synergy of web, NLP and speech technologies and producing new exciting research and 2) fostering a research community for engineering language resources for SDS both via the free data exchange and the release of open-source software for inducing linguistic resources. Furthermore, providing data and linguistic resources for both academic and commercial use will democratize the development of spoken dialogue systems and open it up to a wider audience of developers, as well as, lead to improved technologies for speech services development.

The Speech Service market is a very fragmented market, broken down in two main categories: IVR providers and Speech Application providers. Speech application platform (IVR) providers focus on delivering all the elements necessary for the optimal operation of SDS, and are thus crucial to their success and proliferation, but they expend little effort in addressing the linguistic resources bottleneck and their use in prototyping, in a timely and efficient manner, of new services or the porting of existing ones across languages and domains. This is mainly the task of Speech Application providers that are typically SMEs with limited natural language processing in-house expertise. For these SMEs the lack of multilingual resources (and tools to quickly prototype resources for new domains and languages) is a barrier to entry and penetration of additional markets (especially in Europe).

The SME partners of PortDial have devised their own approaches to commercial exploitation of the outcomes of the project given the niche markets that each partner is active in. VoiceWeb mainly focuses on exploiting PortDial output as a product, enabling rapid and cost-effective porting of voice applications into new application domains and languages. The target segment for this product will be SMEs worldwide in the

mobile application development industry lacking the expertise/resources to develop multilingual speech services in-house.

Expert System exploitation focus is on creating a leading position in the language engineering market, where it already enjoys a key role, providing to system integrators comprehensive solutions that complement and integrate Expert System existing offerings for text processing with tools and linguistic resources for spoken dialog systems. Moreover Expert System is also aiming at commercially exploiting the emerging market of multilingual language resources creation and brokering. The business idea is to define and implement a standardized web service interface for each language resource in order to make them available as web services on a pay per use model basis. Expert System will also include in the above mentioned linguistic resources marketplace the linguistic resources for sentiment analysis coming out from another ongoing project (Eurosentiment: Language Resource Pool for Sentiment Analysis in European Languages).