# Towards a Welsh Language Intelligent Personal Assistant

## A Brief Study of APIs for Spoken Commands, Question and Answer Systems and Text to Speech for the Welsh Government

Stefano Ghazzali

Dewi Bryn Jones

Delyth Prys

October 2015

Language Technologies Unit, Canolfan Bedwyr, Bangor University

http://techiaith.bangor.ac.uk

PRIFYSGOL BANGOR UNIVERSITY

Llywodraeth Cymru
Welsh Government

S4C

# Contents

# 1 Abstract

It is increasingly possible for users to speak in natural English with their devices and computers in order to complete certain tasks, obtain assistance and request information.

This study provides a review of such intelligent personal assistants in use today, their internal architectures in terms of language technology components and opportunities they provide for supporting users with a different preferred language, in our case Welsh.

The study concludes that the architectures of popular and well known intelligent personal assistants are too closed for our purposes, and that their development kits and APIs are too limited to make adoption by researchers and developers for supporting natural Welsh language feasible. Consequently this study recommends how the Welsh Language Communications Project can progress in realising a Welsh language Intelligent Personal Assistant with open architectures and platforms.

# 2 Background

It is increasingly possible for you to speak with devices such as your phone or computer in order to command and control applications and devices as well as to receive intelligent and relevant answers to questions voiced in natural language.

Such capabilities are possible as a consequence of recent advancements in speech recognition, machine translation and natural language processing and understanding. As such they are the prime enablers for a disruptive change and a fundamental shift in how users and consumers engage with their devices and how they more widely use technology.

If looked at in its wider historical context, this is only the next step in the evolution of human computer interaction; from keyboard, to mouse, to touch, to voice and language.

There are four main commercial platforms driving this change, namely Siri, Ok Google, Microsoft Cortana and Amazon Alexa, as well as some lesser known open platforms. To date, these provide their powerful capabilities in English and some other major languages, with little evidence that they are likely to extend their choice of languages to the 'long tail' of smaller languages, including Welsh, in the near future.

This project therefore has been sponsored by the Welsh Government through its Welsh Language Technology and Digital Media Fund and S4C to ensuring that users with a preferred language of Welsh are not left behind in such developments. It will lay the foundations for a range of Welsh language technologies to be used in such environments, including improving the work done to date on Welsh language speech recognition as well as machine translation for leveraging some of capabilities provided via English language based technologies. The project will stimulate the development of new Welsh language software and services that could contribute to the mainstreaming of Welsh in the next phase of human-computer interaction.

## 3 Methodology

We looked initially at the intelligent assistants on offer by the four most active companies, namely; Google, Microsoft, Apple and Amazon. Any associated documentation for developers was researched along with any published APIs and SDKs for assessing any potential for adapting to Welsh language users.

Some basic keyword searches on the web identified a number of open source projects as well as some lesser known on-line API services. These were also evaluated by examining documentation and source code for adaption for Welsh language users.

A generic architecture (as seen in section 4.1) common to all intelligent assistant was realised and used to fully evaluating each platform in terms of feasibility and opportunities for adoption for Welsh.

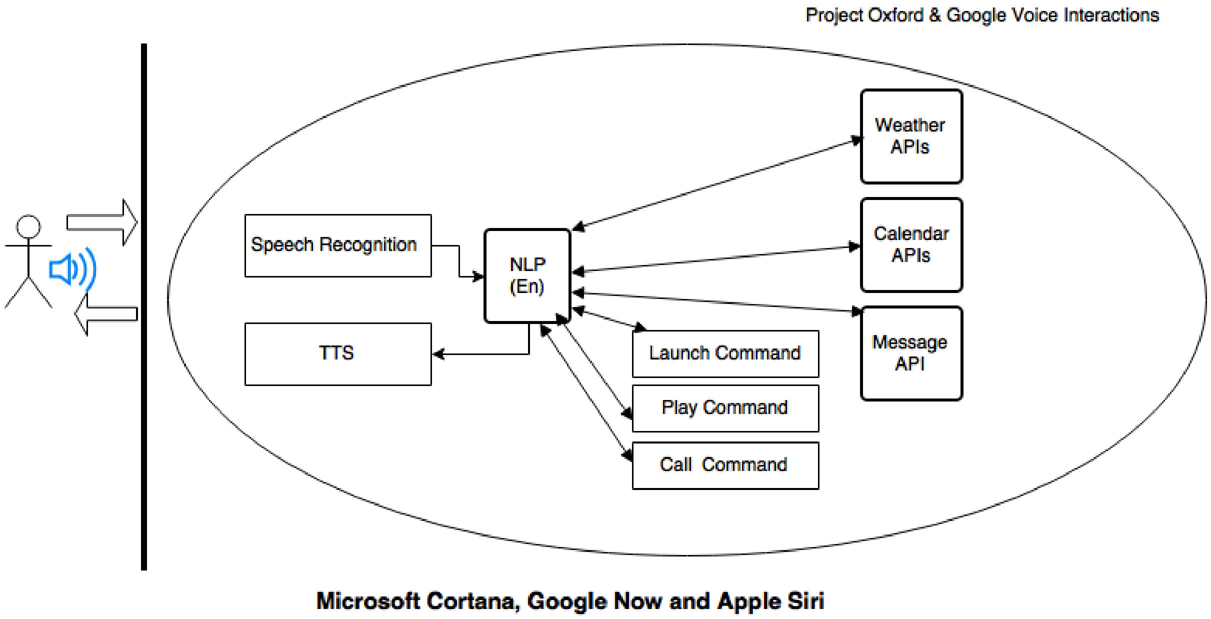
A strategy for fulfilling the objectives of the Welsh Language Communications Project was derived after completion of evaluation exercise.

For the purposes of keeping this study simple, no emphasis was placed on the language in which responses and results are presented in.

# 4 Results

## 4.1 Generic Digital Personal Assistant Architecture

The following figure illustrates the basic architecture that all intelligent assistants share:



A user speaks in natural English language to the device or computer.

1. An English speech recognition engine captures the audio and converts the user's spoken wish into text and is handed over to an English natural language processing (NLP) component.
2. The NLP component is able to process the text and gain an understanding on what the user intends or wants and it thus able to identify which component or API that will fulfil the request (perhaps with the aid of each API or component).
3. The NLP component constructs a message with the various arguments for communicating with a module according to its API specification.
4. In cases where an answer is required, the NLP accepts a response from the fulfilling API and generates a sentence with results included
5. The natural language sentence result is handed to an English text to speech engine for voicing back to the user.

For example, the user may ask "Do I have any meetings today?"

1. The speech recognition engine converts the audio to the text 'DO I HAVE ANY MEETINGS TODAY' and sends it to the NLP component.

2. The NLP component processes the text and recognises it as a question ('DO I HAVE'), recognises 'MEETINGS' and 'TODAY' as keywords associated with

time and a calendar. The NLP polls the calendar API for confirmation, which it agrees.

3. The NLP constructs a message according to the Calendar API specification and sends it. e.g.

```
getEvents(type=meeting, date=today);
```

4. The NLP accepts a list in response, perhaps in a format such as JSON:

```
{"success": true, "events": [ {"date": "2015-09-25",
"time":"10:30am", "name":"Meeting to discuss a new
Welsh language project","location":"Bangor
University, Ogwen Building, Room 234"},{"date":
"2015-09-25", "time":"12:30pm", "name":"Lunch with
Delyth","location":"Terrace Restaurant, Bangor
University"} ]}
```

From which it generates the sentence:

> "Yes you do. At *10:30 this morning* you have a *meeting to discuss a new Welsh language project in Bangor University, Ogwen Building, Room 234*. Then at *12:30* you have *lunch with Delyth in the Terrace Restaurant, Bangor University*"

5. The text to speech engine receives the sentence result, and voices:

> "Yes you do. At half past ten this morning you have a *meeting to discuss a new Welsh language project* in *Bangor University, Ogwen Building, Room two hundred and thirty four*. Then at *half past twelve* you have *lunch with Delyth in the Terrace Restaurant, Bangor University*"

## 4.2 Microsoft Cortana

With the release of Windows 10, Microsoft has released its speech enabled personal assistant app called Cortana. Cortana isn't limited to Windows 10 computers and devices, but is also available for Android and iOS.

Cortana's developers' homepage (http://dev.windows.com/en-us/cortana) states that:

> *"While real-world personal assistants do a lot by themselves, they often rely on a network of experts to help with specific tasks.*
>
> *In much the same way, Cortana knows her users and what they're trying to do, and can use the apps on a user's device as her own network of experts. Make your app part of that network."*

Microsoft provides a means for developers to integrate their apps with Cortana so that they are able to easily provide an intelligent speech interface to their apps. In turn, Cortana's range of capabilities is extended with integrated apps similar in theory to the module APIs as seen in Figure 1.

Microsoft goes further in providing text to speech services and 'Voice Command Definitions' (VCDs) ( simple XML definition files for extending Cortana) that developers can use to build their own natural language speech interactions that start with Cortana.

Cortana's architecture has the speech recognition and natural language processing and understanding encapsulated into one super-component. The architecture does not separate the two functionalities and provide separate API access to the either component, in particular its natural language processing components.

Such a facility would enable a Welsh language intelligent personal assistant to integrate Cortana behind a Welsh language speech recognition engine and machine translation components.

However, all speech interactions via Cortana must start with audio in the languages that Cortana is said to support. It is not feasible to consider Microsoft Cortana further in this project.

### 4.2.1 Microsoft Project Oxford

Microsoft have a range of cloud based language technologies APIs and SDKs hosted at ProjectOxford.ai that serve as the core technologies of Cortana. They provide a means for developers to integrate Microsoft's technologies in more advanced scenarios and solutions than those permitted merely via Cortana.

In particular:

- Speech APIs
  - Speech Recognition – convert spoken audio to text (as full or partial recognition results) in real time.
  - Speech Intent Recognition – in addition to returning text from audio, the API returns structured information on any intent . Your app is able to parse and decide on any further action.
  - Text to Speech – a cloud based text to speech engine.

- Language Understanding Intelligent Service (LUIS) – allows you to use intent recognition without speech recognition. In addition it allows developers to create new intents and model how certain natural language expressions map to structured intent data and parsing in developers' apps.

LUIS does not support forming intent models expressed in Welsh. However, the service could be integrated to process Welsh to English machine translation results.

All APIs are currently available in private beta, available by invitation only. They are not therefore available for further consideration in this project.

## 4.3 Google Voice Actions

Google Voice Actions allow developers to add voice command capabilities to their apps on Android phones, tablets, televisions and Wear devices.

Voiced commands are spoken to Google Search and routed to apps according to phrases that are configured to trigger certain actions. For example, “Ok Google, turn on the lights on MyApp” would cause the MyApp app to open with its API given instructions to set the lights to on. [3]

Google’s Voice Search is able to support simultaneously 50 languages, but not Welsh. Google’s speech recognition for Google Voice Actions, and its ‘OK, Google” service is also expanding its support of languages. Support for Welsh, with no basic support in Voice Search, seems to be a long way off however.

Google Voice Actions offering to developers is similar in construction and scope to that of Microsoft Cortana. Therefore it falls short in its architectures openness and flexibility and cannot be considered further in realising a Welsh language intelligent personal assistant.

## 4.4 Apple’s Siri

Siri is Apple’s personal assistant offering as found in its iOS based devices since a number of years, and has become established in people’s day to day use of their iPhones and iPads.

In contrast to the other companies offering intelligent personal assistant platforms, Apple have not been forthcoming of opening access and sharing Siri technology with developers, much to developers’ frustration and anger. Siri has been entirely closed, sharing the knowledge that they themselves have integrated around 35 APIs to fulfil tasks into Siri.

With the release of iOS 9 and El Capitain however, Apple allows for the deep linking of iOS apps with Search and Siri in much the same approach and scope as Google Voice Actions. iOS 9’s new Search API allows developers to create keywords that Siri’s own speech recognition will recognise for triggering any deep linking activity.

These developments in iOS Siri are still not sufficiently open and flexible in aiding the development of a Welsh personal assistant, and thus Siri cannot be considered further in this project.

## 4.5 Amazon Echo’s Alexa

The new Amazon Echo has  been available in the United States only since July 2015. Users have initially been very impressed by its capabilities with many video demonstrations existing on websites such as YouTube of Alexa being put through its paces and responding helpfully to challenging questions.  There has been to date no announcement by Amazon as to when Echo will be available in Europe and the UK.

Amazon Echo is a cylindrical device that that has an audio speaker, microphones and integrates an intelligent personal voice assistant. There is no display apart from a rotating light that serves as an activity indicator. Echo supports being able to connect to other devices in the home via WeMo, Hue and other Internet of Things (IoT) technologies.

Alexa is the Amazon Echo’s wake word and the name give to the personal voice assistant which, though accessed from Echo, is powered and hosted by Amazon’s cloud infrastructure.

In conjunction with release the device and services for consumers, Amazon also released the Alexa Skills Kit (ASK) and Alexa Voice Service (AVS) as a means for developers to work with Alexa.

Alexa Voice Service allows hardware makers of devices which can connect to the internet, and which have a microphone and speaker, to add the Alexa assistant feature to their devices. Such a service could be integrated into apps on iOS.

The Alexa Skills Set in the meantime allows developers to expand Alexa’s capabilities by defining intents that the Alexa service is able to understand from natural language texts as recognised by Amazon’s speech recognition and then integrating their APIs in order to service their intents.

Alexa to date is only available in American English and the SDKs described above do not allow for development of a Welsh language Alexa based personal assistant.

Amazon do however provide a means and a fund (The Alexa Fund which has $100 million) for researchers and innovative companies to contribute to improving the underlying technologies, namely speech recognition, natural language understanding and text to speech.  The Unit has made initial applications for adding Welsh and multilingual support for Alexa.

Otherwise, Amazon’s Alexa cannot be considered further in this project.

## 4.6 Closed Architectures Conclusions

We have shown that the intelligent digital assistant platforms provided by the four largest and most active companies have closed architectures. This means that access to researchers and developers is limited and focused primarily to those who are able to extend the assistant’s capabilities by integration of their apps and APIs.

Adapting to support speech input and processing for other languages is impossible in all examples.

Consequently it will be impossible to trigger integrated apps and APIs via Welsh language voice commands. Additionally it will be impossible to integrate Welsh language apps and services in order to expand assistants' capabilities to users who require Welsh language specific assistance.

However, this does not preclude the possibility that such large and active companies may extend their choice of languages to include smaller languages such as Welsh through the growing realisation that although markets in the 'long tail' of smaller languages may be individually small, together they provide a sizable proportion of potential users and customers.  The modularization of components and the provision of vital resources for the development of the underlying technologies in freely accessible repositories such as the Welsh National Language Technologies Portal, such as envisaged under the current project, may also encourage take-up by the larger companies.

## 4.7 Open Alternatives

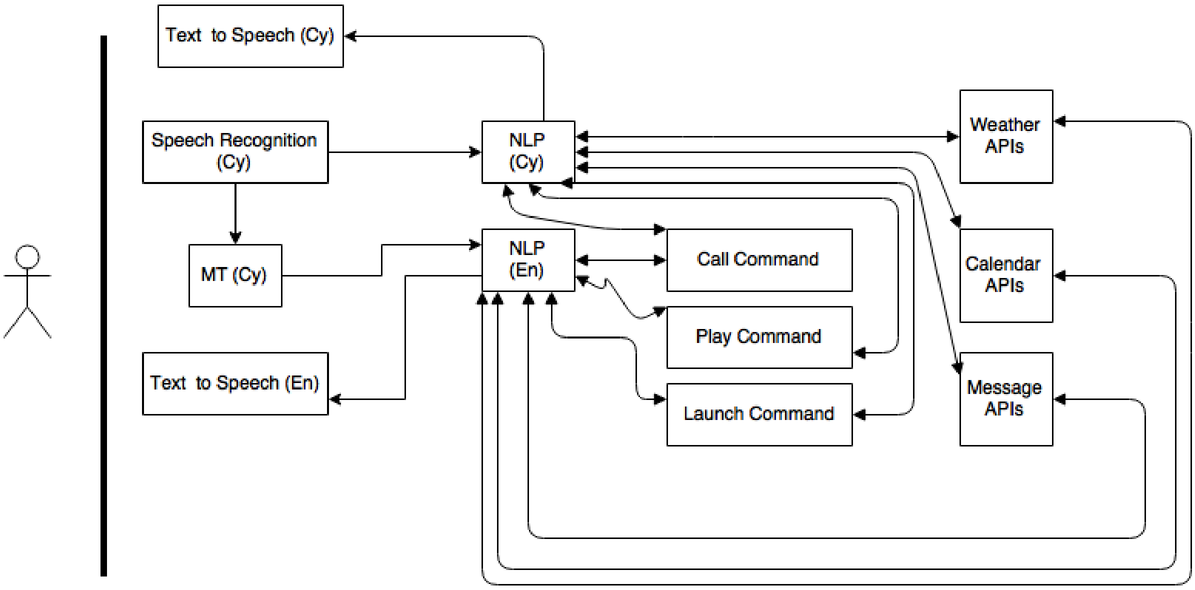
A Welsh language personal digital assistant requires architectures that are far more open and flexible in which a number of language technologies can be integrated.

A number of other intelligent personal assistant platforms exist that are more open in terms of architectures; APIs and source code were evaluated as the prime source for a solution.

Our set of criteria was:

- a Welsh language speech recognition engine could be integrated.
- The NLP for language understanding could either be adapted or replaced
- Responses could be provided via Welsh language text to speech.
- APIs and modules that fulfil tasks but which are based in English language usage, could still be included with novel integration of Welsh to English machine translation.

The following figure illustrates a desirable architecture:



## 4.7.1 Jasper

Project Jasper is a very simple open source project primarily developed to run on an internet connected Raspberry Pi with only a microphone and loudspeaker attached [8]. Its implementation provides the general architecture as seen in the figure in section 4.1.

Components that deliver a specific functionality can be easily integrated. Thus, it allows the developer to select which speech recognition with integration code already existing for speech recognitions on the device such as PocketSphinx and Julius/HTK or in the cloud such those provided by Google. Simple alterations to the code can be made to include other components such as machine translation.

A ‘brain’ component delivers a means for easily extending Jasper’s capabilities in which it and integrated modules or APIs determine and understand intents.

Text to speech components can also be added or replaced although Jasper already integrates with the Festival Speech Synthesis Engine, which the Unit has experience with and a Welsh voice available.

Thus Jasper is attractive in being a very simple architecture for developing and proof of concept work associated with developing the underlying language technologies components. It is also attractive for use in an education setting and code clubs that would develop and extend capabilities.

### 4.7.2 SIRIUS

Sirius is an open-source end-to-end standalone speech and vision based intelligent personal assistant (IPA). This project is based out of the Clarity Lab at the University of Michigan. It is not connected or associated in any way with Apple's Siri.

Sirius receives queries in the form of speech or images and returns results in the form of natural language. Sirius implements the core functionalities of an IPA including speech recognition, image matching, natural language processing and a question-and-answer system. [9]

Sirius is designed for deployment onto more powerful machines or servers with example client programs available for accessing over the web. Sirius already has code for integrating CMU Sphinx speech recognition.

As the architecture is open and sufficiently modular, Sirius is worth utilizing as a solution that could permit Welsh language IPAs in the cloud accessible from any OS, device and/or wearable.

### 4.7.3 Wit.ai

Wit.ai is a Facebook company charged with building AI platforms that help developers to create apps that understand human language.

Wit.ai's approach however is different from other commercial ventures in that its API services are more modular. Wit.ai claims that they engage with communities by sharing everything that it learns with developers.

Wit.ai's cloud based APIs allows developers to send either the audio or text of a user's request and returns from its natural language components actionable data that can be used in turn to fulfill the task.

Wit.ai is of interest therefore in that its API can accept a Welsh to English machine translation of a request in voiced originally in Welsh.

### 4.7.4 Mycroft.ai

Mycroft.ai is an open source project and product but is still a commercial venture. It is similar to Jasper in that it is based on the Raspberry Pi but has a casing with simple LEDs lights arranged to mimic a face. Its ambition is 'A.I. for everyone' so contains more complex natural language processing. In short, a more open and extensible system than Amazon's Echo. Mycroft supports a number of IoT protocols including NEST, WEMO, Hue and Iris.

As yet, due to ongoing campaigns on Indiegogo, Mycroft.ai product is available only as a pre-order and the source code scheduled to be released after April 1st 2016. Unfortunately it will not be of any use to the current project.

# 5. Conclusions

Only the open alternatives can provide this project with assembling a Welsh language intelligent personal assistant.

The open alternatives that we have evaluated vary in sophistication and complexity and in the support for request they can support.

Jasper is very simple for use in Raspberry Pis whereas Wit.ai and Sirius are cloud based deployment that allow use by many users from multiple sources.

We are fortunate therefore that the open alternatives would be able to support an incremental development strategy, where Jasper can be used for developing a system with one or two limited domain capabilities such as requesting the time in Welsh, and Sirius for more complex requests that require open ended question and answering requests, plus access via apps and/or wearables.

Such an incremental approach allows for some of the Unit's other resources to be introduced into the development process in a controlled manner. For example, the Unit already has a basic Welsh language text to speech voice, which is available for producing speech either from a computer (such as a Raspberry Pi) or via the cloud from the Unit's Cloud APIs service.

We will also be able to develop the speech recognition to support increasing number of possible users request, until, hopefully, a large vocabulary continuous speech recognition engine is realized for accepting any general questions. Already the Unit owns many relevant resources, which it has made available through the Welsh National Language Technologies Portal, that are essential 'building blocks' for developing an intelligent personal assistant for Welsh. These, and new ones created specifically in the course of this project, will be used to build prototypes and proof of concept models using open architecture. They may also be used by others, including companies who keep their own source code closed, to extend their offerings to include Welsh.

This open model therefore offers a way forward not only to Welsh, but also to other smaller language communities, to create an inclusive ecosystem, where both social and economic benefit will be derived from sharing basic resources, where speech enabled devices will empower speakers of all languages to engage with the new generation of digital media and assistants.

# 6. References

1. Nikola Metulev (2015) *Meet the Speech Platform in Windows 10,* Available at: http://metulev.com/meet-the-speech-platform-in-windows-10/ (Accessed: 21 September 2015).
2. Microsoft Project Oxford (2015) *Microsoft Project Oxford,* Available at: https://www.projectoxford.ai/luis (Accessed: 21 September 2015).
3. Laurence Moroney (2015) *Coffee with a Googler: Learn about Google Voice Actions,* Available at: http://googledevelopers.blogspot.co.uk/2015/09/learn-about-google-voice-actions.html (Accessed: 21 September 2015).
4. Google (2015) *Google Voice Interactions - Overview,* Available at: https://developers.google.com/voice-actions/interaction/ (Accessed: 21 September 2015).
5. ISpeech (2015) *ISpeech API Developer Guide,* Available at: http://www.ispeech.org/api (Accessed: 21 September 2015).
6. Politepix UG (2015) *Open Ears Website,* Available at: http://www.politepix.com/openears/ (Accessed: 21 September 2015).
7. Expect Labs (2015) *Getting Started with the iOS SDK,* Available at: https://expectlabs.com/docs/sdks/ios/gettingStarted (Accessed: 21 September 2015).
8. Jasper Project (2015) *Jasper Project Documentation,* Available at: http://jasperproject.github.io/documentation/ (Accessed: 21 September 2015).
9. Johann Hauswald, Michael A. Laurenzano, Yunqi Zhang, Cheng Li, Austin Rovinski, Arjun Khurana, Ron Dreslinski, Trevor Mudge, Vinicius Petrucci, Lingjia Tang, and Jason Mars. Sirius: An Open End-to-End Voice and Vision Personal Assistant and Its Implications for Future Warehouse Scale Computers. In *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, ASPLOS ’15, New York, NY, USA, 2015. ACM. Acceptance Rate: 17%.