Faking Intelligent CALL: the Irish context and the road ahead

Neasa Ní Chiaráin and **Ailbhe Ní Chasaide**
The Phonetics and Speech Lab.
Centre for Language and Communication Studies
Trinity College, Dublin
neasa.nichiarain@tcd.ie

Abstract

Speech-enabled dialogue systems developed within an iCALL framework offer a potentially powerful tool for dealing with the challenges of teaching/learning an endangered language where learners have limited access to native speaker models of the language and limited exposure to the language in a truly communicative setting. This paper explores the major potential of virtual conversational agent systems with inbuilt simulated 'intelligence' for the Irish (endangered) language context.

1 Introduction

Multimodal dialogue systems with inbuilt simulated 'intelligence' have huge potential in language learning/teaching environments. In the context of the many minority/endangered languages, such as Irish (Gaelic), these systems could make even more of an impact. Major difficulties exist for language learners related to the lack of exposure to native speaker models and creating virtual 'native speakers' to converse with learners opens new paths towards overcoming these issues in the socio-linguistic context of endangered languages. Many languages, and particularly minority, endangered and under-resourced languages lack several of the linguistic and technological prerequisites for the construction of 'intelligent' dialogue partners. Nonetheless, as will be illustrated here, interim solutions are possible, which can offer partial dialogue systems, which can still have impact in the teaching/learning context.

Following a brief discussion of the socio-linguistic context of current developments for Irish, this paper presents (i) a simulated intelligent dialogue partner, constructed for Irish language tuition, using synthetic voices and an animated avatar (a talking monkey), (ii) a discussion on how, in the absence of NLP-based resources (yet to be developed for Irish), specific strategies are adopted which allow the impression of 'intelligent' discourse with an agent, and (iii) an outline of the steps envisaged to allow a fuller, more 'intelligent' system, using NLP resources.

2 The socio-linguistic context of Irish language teaching/learning

Irish is the first official language of the Republic of Ireland and is a working language of the European Union. Yet, it is an endangered language (Moseley, 2010) in that it has no monolingual speakers and there are few, if any, domains where Irish is the sole acceptable language. Irish is a compulsory subject of study for all pupils attending second level schools in the Republic of Ireland. Teachers, however, are often second language learners and therefore there is huge variation in levels of proficiency ranging from relatively low communicative competence to traditional native speakers. At second level the recommended annual taught time for Irish is 110 contact hours per year (Eurydice, 2013, p. 10) which means learners lack sufficient input: far more exposure to the language than what is currently available within school hours is need-

Neasa Ní Chiaráin and Ailbhe Ní Chasaide 2016. Faking Intelligent CALL: The Irish context and the road ahead. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC 2016*. Linköping Electronic Conference Proceedings 130: 60–65.

ed. The use of interactive language learning technology in schools is extremely limited and the use of antiquated and dull teaching materials (and sometimes methods) adds further to low levels of motivation.

Since motivation is generally accepted as being the prime factor associated with successful language learning (Robichaud, 2014), the development of virtual world platforms where the learner can interact with an artificial interlocutor/dialogue partner and create the semblance of a natural conversation seems appropriate. The learner can become engaged with the target language and use it to complete specific tasks or engage in games. Though the development of such platforms is still in its infancy, the concept would seem to have a particular attraction in the case of minority or endangered languages.

3 A provisional interim dialogue partner

In the major world languages much effort has been put into creating speech activities which allow learners to engage in spoken interaction with a conversational partner, the most difficult competence for a learner to acquire independently. An initial attempt at providing opportunity for students of Irish to practice conversation is presented here as *Taidhgín*, (pronounced: [tˠ aɪ ɟ iː nʲ]), an 'intelligent' dialogue partner in the form of an animated, smartly dressed monkey. *Taidhgín* was built using Artificial Intelligence Markup Language (AIML), an XML-based open-source programming language which was developed by Richard Wallace and the Alicebot free software community during the period 1995-2000. *Taidhgín* is hosted and run from Pandorabots which is a 'free open-source-based community web service which enables you to develop and publish chatbots on the web' (pandorabots.com). *Taidhgín* has integrated Irish language synthetic voices which are developed as part of the ABAIR initiative (www.abair.ie) in Trinity College, Dublin. Ideally, the chatbot presented here would form part of an end-to-end spoken dialogue system with speech input and output but as there is not yet an automatic speech recognition system for the Irish language, the user must input speech to the *Taidhgín* system by typing into a text box.

Figure 1: *Taidhgín:* the prototype dialogue partner.

Evaluations of *Taidhgín* were carried out nationwide in 13 schools by 228 pupils. The evaluations consisted of (1) eliciting learners' opinions of the overall chatbot platform as a learning environment and (2) evaluating the intelligibility, quality, and attractiveness of the ABAIR text-to-speech synthetic voices used in this platform. Results were very positive to both the learning platform and to the synthetic voices, evidenced by an evaluation by 228 16-17 year old learners of Irish, 73% of whom rated 'intelligibility' at points 4 or 5 (positive or very positive) on a Likert scale; 73% rated same for 'quality'; and 53% rated same for 'attractiveness'. This demonstrates that even a partially 'intelligent' system which exploits speech and language technologies stands to have immediate impact in the Irish educational context. For a fuller account of evaluations see Ní Chiaráin & Ní Chasaide (2016). Further evaluations were carried out on proficient speakers of Irish who are teachers and results were also found to yield similarly high ratings (Ní Chiaráin, 2014).

The Pandorabots system presented here is based on pattern matching whereby all likely responses to *Taidhgín*'s questions are hardcoded. Therefore much content development work was needed in order to give a certain appearance of intelligence to *Taidhgín*, as the system began with no initial Irish language content. The most common errors (grammatical and orthographic) made by Irish Leaving Certificate students (pre-University examinations) have been documented in work by Ó Baoill (1981) and this information was used in the development of *Taidhgín* to build an internal correction system. Currently the most commonly made errors are hardcoded into the system: when learner input is matched to these errors *Taidhgín*

Version used by *Taidhgín* chatbot	**Translation**
Human: Tá **dhá** deartháir agam. ***Taidhgín*:** Ó, tuigim – **beirt** deartháir! Níl aon deirfiúr agat, buachaillí ar fad atá sa teach leat! Agus, an bhfuil na deartháireacha seo níos óige nó níos sine ná tusa?	***Human:*** *I have* ****two*** *brothers.* ***Taidhgín:*** *Oh, I understand –* ***two*** *brothers! You've no sister, all boys in the house! And are these brothers younger or older than you?*

Figure 2: *Taidhgín* feedback: reformulating learner input and recasting the corrected version

reprises a correct version as part of his response. This manner of correction avoids a break in the flow of conversation, which explicit correction would entail. An example of this is presented in Figure 2 and discussed further below.

In addition to this recasting correction mechanism, the log files are made available to the learner and tutor for later review. The grammar and spelling checkers which are available in Firefox are also used so that errors in the input are highlighted in the learners' text box, allowing correction of the text before submission. Given the complex orthography of Irish this ensures that the users' spelling errors don't result in a breakdown of the communication.

At the present stage of development 11 topics (aligned to the second level oral examination curriculum, including 'family', 'holidays', 'hobbies', etc.) consisting of 3,670 categories have been added in order to make *Taidhgín* seem 'intelligent' (category = a conversational turn consisting of a question with potentially multiple responses, including anticipated errors, as discussed above).

Early elements of grammar and spelling correction facilities have been included in the prototype design to date. The example in Figure 2 illustrates one example, i.e. the numerical system in Irish, which is relatively complex. The learner's error and *Taidhgín*'s corrected versions are shown in boldface. The number '2', for example, can be expressed as *dó*, *dhá*, *beirt*, *dara*, *dóú* / *dhó* depending on the context in which it arises. For example, the terms *dhá* and *beirt* are identically used but qualify different types of nouns: *dhá* is used for inanimate objects (e.g. *dhá chupán* 'two cups') while *beirt* is used for humans (e.g. *beirt chailín* 'two girls'). Both correct and incorrect usages are anticipated in the preparation of the categories for *Taidhgín*. If the learner used ****dhá*** *dearthár* *'two brothers' instead of ***beirt*** *dearthár* 'two brothers' the correct version is recast by *Taidhgín* and the conversation continues.

Another area with which learners tend to have trouble concerns the two forms of the verb '**to be**' in Irish. The copula ***is*** exhibits a characteristic of permanency and stability, and is used to express nationality or profession, for example, ***Is*** *múinteoir mé* ➔ 'I **am** a teacher'. The substantive verb ***bí*** (*tá / níl* 'I am / I am not') is employed to describe a more transitory state (***Tá*** *mé ag obair* ➔ 'I **am** working'). Again, in the AIML categories common errors that learners make were predicted and hardcoded so that the system could provide corrective feedback as appropriate.

4 Next steps towards incorporation of 'intelligence'

In its current implementation, *Taidhgín* is faking it. He is not intelligent in the sense of being able to identify an error and correct it: rather, he simply has hardcoded error versions for very specific sentences pertaining to the topics developed so far. Our vision for the future is to give *Taidhgín* more of a brain, so, rather than merely pattern matching, the system can access correct/incorrect usage of grammatical rules, etc. and formulate correct versions.

As part of a Digital Plan for Irish Speech and Language Technology (2016 - 2026), commissioned by the Department of Arts, Heritage Regional, Rural and Gaeltacht Affairs, NLP and speech technology resources are being developed for Irish and we look to some of these developments to grow *Taidhgín*'s intelligence. Resources that are already available include a grammar

checking engine, (Scannell, 2005), available in Firefox and usable with *Taidhgín*, a morphological analyser (Uí Dhonnchadha, Nic Pháidín, & Van Genabith, 2003), a part-of-speech tagger (Uí Dhonnchadha & Van Genabith, 2006), and a chunker (Uí Dhonnchadha & Van Genabith, 2010).

A recently developed resource is the semantic WordNet for Irish (O'Regan, Scannell, & Uí Dhonnchadha, 2016) which classifies lexical units into categories, e.g. *profession* or *nationality*. If *Taidhgín* can detect such information in the learner's input it should enable him to spot whether the correct form of the verb 'to be' is being used, i.e. the copula ***is*** or the substantive verb ***bí***. This is a simple case where error detection can be generalised rather than being dependent on hardcoding.

Corrective feedback for the learner can be presented either implicitly (where the correction is recast by the dialogue partner and flow is not interrupted) as illustrated in Figure 2, or explicitly (more on this below).

As with the forms of the verb 'to be', animate and inanimate nouns can be classified in WordNet and this can be used to identify correct/incorrect usage of the numerals, as discussed above (see also Figure 2).

The use of NLP tools will serve different purposes in making *Taidhgín* a useful pedagogical aid. For example, the morphological analyser and generator (Uí Dhonnchadha et al., 2003) can be used both for the creation of CALL content (quizzes, etc. for grammatical drilling) and to allow *Taidhgín* to identify if the learner's input violates grammatical rules such as tense and verb conjugation, etc. Similarly, the spelling and grammar checker/corrector (Scannell, 2005) can be used for developing drills as well as ensuring comprehensible learner input to *Taidhgín* so that the system can recognise the learner's string and respond appropriately, ensuring there are fewer breakdowns in communication.

The future plan is to incorporate these new technologies into the *Taidhgín* conversational pedagogical agent platform in order to develop a combination of form-focused instruction and meaning-focused conversation.

It is intended that the learner would start by *chatting* to *Taidhgín* and if/when errors should be detected by the system, learners would be given the option either to leave the conversation, focus on form and concentrate on a specific aspect of the language with which they have difficulty (see Figure 3: *Trialacha Taidhgín* 'Exercises with Taidhgín' for options to train certain linguistic features) or to continue with meaning-focused conversation, maximising 'flow' or the engagement of the learner with the task (Csíkszentmihályi, 1988) while the learning process is being steered with the inclusion of appropriately scaffolded material.

Up to now we've talked about how *Taidhgín* might detect errors. It is important to note that there are several remedial approaches that can be taken beyond the implicit recasting and explicit focus on form drilling mentioned above. If errors are logged and it transpires a particular error is made a set number of times, *Taidhgín* could adapt the dialogue so that areas where the learner needs additional practice are foregrounded (implicitly, e.g. if past tense formation is a problem, *Taidhgín* could frame his questions in the past tense).

Alternatively, *Taidhgín* can explicitly draw the learner's attention to the correct form with 'did you mean X?' questions. *Taidhgín* could even prompt the learner to leave the guided free dialogue and spend some time instead practicing using a fun, contextualised exercise designed specifically to drill a particular linguistic feature of Irish. The interface to such drills is illustrated in Figure 3.

Personal profiles will be constructed for individual language learners so that the responses by the avatar may be more finely tuned to the individual. This not only helps a more adaptive learning environment but should enable a degree of personalisation of content in such a way as to engage the learner by having the avatar establish a rapport with them.

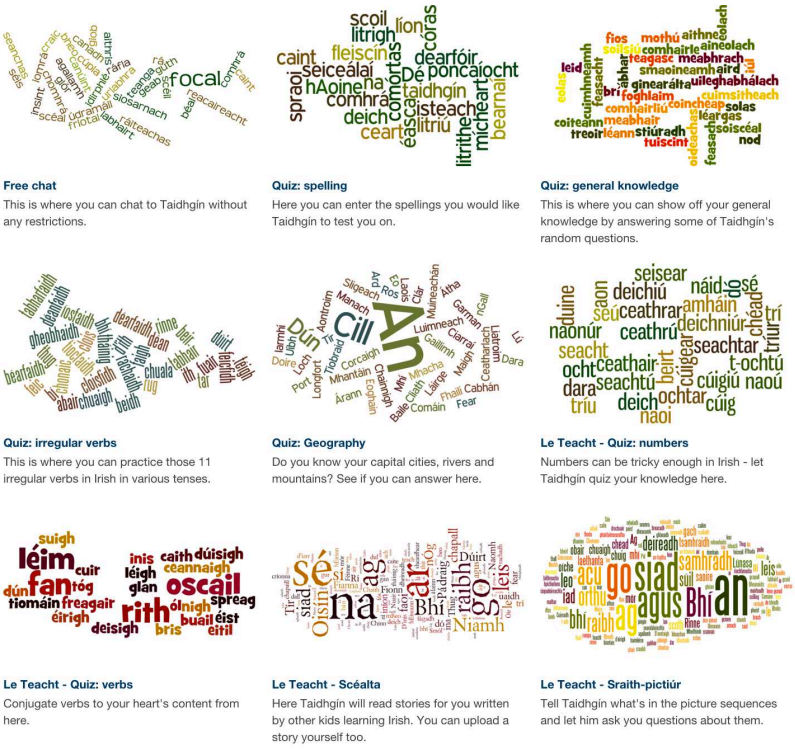

Figure 3: *Trialacha Taidhgín* 'Exercises with *Taidhgín*': interface to a range of focus on form exercises including, for example, quizzes on irregular verbs, spelling, general knowledge, etc

This paper has discussed those aspects of intelligence that we would hope to work towards incorporating into such dialogue systems. Of course there are other aspects of *Taidhgín's* growing brain that will need some attention: he will need to be able to 'hear' what the learner says to him in order to conduct a more meaningful conversation. Within the context of the Digital Plan for Irish (2016-2026), speech recognition is envisaged. Incorporating recognition into *Taidhgín* will enable a full end-to-end spoken dialogue system. A full recognition system will inevitably take time to develop but even a partial system could, in the short-term, provide interesting options. It will be important to ensure that the future spoken output of *Taidhgín* can handle the conversational prosody of true dialogues.

5 Conclusions

The overall goal is to harness the emerging technologies in a way that will enable more effective language learning. It is planned to incorporate more NLP resources as well as speech resources into the current prototype of the *Taidhgín* system which will both ensure that the flow of dialogue is less likely to fail, and also enable the dialogue system to pick up on incorrect forms, respond appropriately to the learner and provide intelligent corrective feedback.

As the simple prototype illustrated above indicates there is great potential for developments in this field. It is hoped that the *Taidhgín* prototype might benefit those dealing with the ever more daunting task of maintaining endangered languages through education. The future survival of Irish and many such endangered languages will depend on how effectively they can be transmitted to the next generation. In this context, there is some urgency with ensuring that our educational resources make full use of what modern speech and language technologies have to offer.

Acknowledgments

The authors gratefully acknowledge funding from the Department of Arts, Heritage, Regional, Rural and Gaeltacht Affairs, Ireland (the ABAIR project) and from An Chomhairle um Oideachas Gaeltachta & Gaelscolaíochta (the CabairE project).

References

Csíkszentmihályi, M. (1988). The flow experience and its significance for human psychology. In M. Csíkszentmihalyi & I. S. Csíkszentmihalyi (Eds.), *Optimal experience: Psychological studies of flow in consciousness* (pp. 15–35). Cambridge: Cambridge University Press.

Eurydice. (2013). *Recommended Annual Taught Time in Full-time Compulsory Education in Europe 2012/13*. *Eurydice - Facts and Figures, European Commission*. Retrieved from http://eacea.ec.europa.eu/education/eurydice/documents/facts_and_figures/taught_time_EN.pdf

Moseley, C. (Ed.). (2010). *Atlas of the world's languages in danger* (3rd ed.). Paris: UNESCO Publishing. Retrieved from http://www.unesco.org/culture/en/endangeredlanguages/atlas

Ní Chiaráin, N. (2014). *Text-to-Speech Synthesis in Computer-Assisted Language Learning for Irish: Development and Evaluation*. (Unpublished Doctoral thesis, CLCS, Trinity College, Dublin).

Ní Chiaráin, N., & Ní Chasaide, A. (2016). Chatbot Technology with Synthetic Voices in the Acquisition of an Endangered Language: Motivation, Development and Evaluation of a Platform for Irish. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. 23-28 May 2016 (pp. 3429-3435). Portorož, Slovenia: European Language Resources Association (ELRA).

O'Regan, J., Scannell, K., & Uí Dhonnchadha, E. (2016). lemonGAWN: WordNet Gaeilge as Linked Data. In *LDL 2016 – 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources* (pp. 36–40).

Ó Baoill, D. (1981). *Earráidí scríofa Gaeilge. Cuid 3, réamhfhocail agus comhréir : earráidí a tharla in aistí Gaeilge na hÁrdteistiméireachta, 1975*. Baile Átha Cliath: Institiúd Teangeolaíochta Éireann.

Robichaud, A. (2014). Interview with Noam Chomsky on Education. *Radical Pedagogy*, *11*(1). Retrieved from http://www.radicalpedagogy.org/radicalpedagogy.org/Interview_with_Noam_Chomsky_on_Education.html

Scannell, K. (2005). An Gramadóir. Retrieved October 3, 2016, from http://borel.slu.edu/gramadoir/

Uí Dhonnchadha, E., Nic Pháidín, C., & Van Genabith, J. (2003). Design, implementation and evaluation of an inflectional morphology finite state transducer for Irish. *Machine Translation*, *18*, 173–193. http://doi.org/10.1007/s10590-004-2480-9

Uí Dhonnchadha, E., & Van Genabith, J. (2006). A part-of-speech tagger for Irish using finite-state morphology and constraint grammar disambiguation. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*.

Uí Dhonnchadha, E., & Van Genabith, J. (2010). Partial dependency parsing for Irish. In *7th International Conference on Language Resources and Evaluation (LREC 2010)*.