Modelling Multiple Language Learning in a Developmental Cognitive Architecture

Ioanna Giorgi[1a], Bruno Golosio[2], Massimo Esposito[3], Angelo Cangelosi[1b], Giovanni L Masala[4]

[1]The University of Manchester, UK, [a]ioanna.giorgi@manchester.ac.uk, [b]angelo.cangelosi@manchester.ac.uk
[2]University of Cagliari and INFN, Italy, golosio@unica.it
[3]Institute for High Performance Computing and Networking - National Research Council, Naples, Italy, massimo.esposito@icar.cnr.it
[4]Manchester Metropolitan University, UK, G.Masala@mmu.ac.uk

Abstract. In this work, we model multiple natural language learning in a developmental neuroscience-inspired architecture. The ANNABELL model (*Artificial Neural Network with Adaptive Behaviour Exploited for Language Learning*), is a large-scale neural network, however, unlike most deep learning methods that solve natural language processing (NLP) tasks, it does not represent an empirical engineering solution for specific NLP problems; rather, its organisation complies with findings from cognitive neuroscience, particularly the multi-compartment working memory models. The system is appropriately trained to understand the level of cognitive development required for language acquisition and the robustness achieved in learning simultaneously four languages, using a corpus of text-based exchanges of developmental complexity. The selected languages, Greek, Italian and Albanian, besides English, differ significantly in structure and complexity. Initially, the system was validated in each language alone and was then compared with the open-ended cumulative training, in which languages are learned jointly, prior to querying with random language at random order. We aimed to assess if the model could learn the languages together to the same degree of skill as learning each apart. Moreover, we explored the generalisation skill in multilingual context questions and the ability to elaborate a short text of preschool literature. We verified if the system could follow a dialogue coherently and cohesively, keeping track of its previous answers and recalling them in subsequent queries. The results show that the architecture developed broad language processing functionalities, with satisfactory performances in each language trained singularly, maintaining high accuracies when they are acquired cumulatively.

Keywords: neural network, cognitive system, natural language understanding, multilingual system.

1. Introduction

Artificial intelligence has led to impressive achievements on a variety of complex cognitive tasks, matching or beating humans. This includes playing Go or videogames [1], speech and NLP [2], object and facial recognition [3]. With reference to natural language processing, the use of deep learning models in the last five years has strongly propelled it forward, with considerable advances in real-world NLP applications, like image captioning [4,5], visual question answering [6,7], web search and information retrieval [8,9], sentiment analysis [10,11] and recommender systems [12,13]. Architectures inspired by human cognition have been proposed to model language comprehension, learning and reasoning. They attempt to integrate neural models of language into comprehensive cognitive architectures compatible with current knowledge on how storing and processing of the verbal information occurs in the brain. However, natural language development and understanding is arguably one of the least understood human capabilities from a cognitive perspective. One reason could be the complexity of human language and the concurrence of general mechanisms of information processing in the brain's architecture [14]. Developing humanlike cognitive systems able to acquire one or more languages, analyse them into parts, comprehend spoken or written language, and produce natural-sounding sentences, is yet a significant open problem.

In this work, we present the cognitive system ANNABELL [32,42], endowed with the capability of processing and producing four natural languages, with significant differences and complexity levels, with the final aim to contribute to the computational understanding of appropriate characteristics that favour multi-language development and understanding. We propose a general solution for learning the languages, where the system architecture and procedural knowledge used in language elaboration remains the same i.e. the ability to control the flow of information among different buffers and memory systems of the model (see section 3.1). We claim that many aspects of language development and language processing skills can be described in terms of working memory [47] operations. Implicitly, we also claim that there is a level of language processing that involves the flow of information among working memory buffers, which is language independent [37]. To fulfil our aim and prove these research claims, we perform an extensive experimental training and validation, with four different languages with peculiar lexical, syntactical, morphological, organisational, and semantic aspects. In particular, a cumulative approach is adopted to assess the system's capacity of generalisation of multiple languages at scale, under increased language complexity, and to prove that the system can successfully disambiguate the languages, whilst delivering appropriate conversation, for several tasks. The term cumulative here refers to simultaneous training of the languages, in which all datasets are learned jointly before the test. We were inspired by the studies on multilingualism [46], to propose a cognitively plausible model that acquires language capabilities of multiple natural languages simultaneously. The cumulative training can be useful to understand the cognitive processes involved in simultaneous language acquisition and how the brain becomes "tuned" to whatever languages experienced since birth. It is motivated by studies whose findings show that the activity of the bilingual brain reflects the languages it has been exposed to [50]. From a non-cognitive perspective, the growth of multilingual speakers, outnumbering their monolingual counterparts, has led to greater commercial interest in multilingual services and

households. In our work, we emphasise the concept of "acquiring" as opposed to "learning", as the latter refers to the process of studying a language and, how linguistic forms (grammar, semantics and phonology) interact with one another [41]. Language acquisition best describes the type of training performed in ANNABEL that is in the form of communicative activities, in which the system experiences language use and reproduces it closely to communicate back.

To assess the system's degree of skill in multiple language acquisition, we first trained it in all four languages separately and measured the performances (accuracies) for each language independently. The two approaches (individual and cumulative) are independent of one another and are carried out to compare the behaviour of the system when it acquires a language alone or jointly with other languages. Finally, we explored the (multilingual) competences of the system in handling context questions, on a short narrative of preschool literature, in particular, the ability to comprehend the text and dialogue with the human coherently and cohesively, logically linking past answers to subsequent questioning.

The paper is organised as follows. Section 3 describes the methods used in this work, starting with an overview of the model and the dataset and further motivating the selection of the languages. Section 4 presents the experimentation carried out using an individual cross-validation, a cumulative (and incremental) cross-validation approach, respectively. Section 5 describes the set up and results on narrative text comprehension. The paper is concluded in Section 6.

2. Related Work

The successes in language-related AIs have been facilitated by the scale-up of already existing neural network models, i.e. convolutional neural networks and recurrent neural networks that have only recently produced significant achievements over state-of-the-art NLP systems, due to the thriving availability of huge databases and substantial amounts of computing power. The large distribution of machine learning programming frameworks, the open sourcing of datasets and the pre-trained state-of-the-art systems that can be downloaded and tested on new textual inputs [14], have greatly influenced the growth of research in NLP. This is reflected in a great commercial interest in the deployment of human language technology, in the form of conversational systems for personal mobile phones (e.g. Siri and Google Assistant) or embedded in standalone devices (e.g. Amazon Echo and Google Home). However, these are trained to respond only to a pre-set number of requests for specific use cases. Google Assistant has recently launched multilingual support, allowing for simultaneous bilingual use and jumping between languages across queries, without hitting the language settings [15]. However, to reduce the increased processing costs and unnecessary latency that derive from the more sophisticated architecture of the system required for each new language, Google Assistant uses an early identification approach to make a quick switch to single monolingual recogniser, to manage multilingual queries. In addition, it has greatly limited the list of candidate languages the user can choose (no more than two at a time). Another open platform is IBM Watson, equipped with advanced natural language abilities, combining NLP and machine learning, to understand the structure of nine different languages, i.e. to parse - to identify verb, nouns, adjectives and other parts of speech [16]. Despite, it has been criticised as a "finicky eater" to data enterprises, fussy about the users' requirements and very demanding on data preparation standards [53]. Moreover, although trained on a large corpus in nine different languages, it cannot ingest languages altogether and converse them simultaneously [16].

Despite the remarkable success of deep learning in different NLP tasks, either monolingual or multilingual, there remain yet significant challenges. Indeed, the current deep learning methods have been scaled up and improved, but, in turn, they have augmented their complexity. Therefore, even if originally inspired by the complex hierarchical organisation of the cerebral cortex, they have assumed the form of empirical engineering solutions to solve specific NLP problems. Moreover, they are extremely data hungry. Unlike humans who are far more efficient in learning complex rules from a few examples, they have shown to work best only with thousands, millions or even billions of training examples [17,18]. Their biggest drawback is the inability to explain their outputs, which is relevant for natural language systems when they are asked to explain the process followed for comprehending text, especially if a decision must be taken (e.g., booking a hotel room by speaking to a conversational assistant) [19]. Thus, current deep learning systems fail to provide a human-like, computational model of cognition able to provide intelligible insights about human brain mechanisms on how one or more languages are acquired, comprehended and produced [14, 20].

We have narrowed our focus mainly on neural systems that model brain processes. While current NLP systems are undoubtedly impressive, the strategies they employ differ greatly from those humans use for language acquisition [21]. The most pertinent contributions in this field date early back, starting with a cognitive neural architecture [22-24], with capabilities to parse script-based stories, store them, generate paraphrases of the narratives, and answer questions, even though its effectiveness has been proved only on a very small corpus. Neoteric works [25, 26] include a neural model of brain areas involved in language processing, able to learn grammatical constructions and generalise to novel constructions and to the production of sentences [27]. A bilingual version of this model has shown to generalise strictly on grammatical constructions [28], with application to 15 languages [29], on human-robot interaction corpora. The computational model of (first) language acquisition in [30] demonstrates language construction from scratch by combining bottom-up and top-down learning processes. An embodied cognitive architecture based on a biologically-grounded theory of the brain and mind, has attempted to solve the symbol grounding problem and acquire language capabilities, by generalising on narrative constructions within a robotic architecture [31]. The cognitive architecture used here, ANNABELL (Artificial Neural Network with Adaptive Behaviour Exploited for Language Learning) [32] has proven able to develop a broad range of functionalities for elaborating verbal information. It is based on a very large-

scale neural network, originally intended to help understand the cognitive processes in early language development [32].

There exists a considerable body of literature on multiple language acquisition and understanding. One method employed by [33] simulates bilingual lexical representations and interactions with an unsupervised SOM network, using large-scale linguistic data from children's early lexicons (CHILDES). Another unsupervised learning algorithm has been presented by [34] to simulate L2 construction learning from bilingual input, without modelling the cognitive behaviour of how humans learn a second language (L2). The method proposed by [35, 36], presents a CNN model to learn common multilingual representations and image descriptions, using images as a pivot, to improve image understanding and search. However, we argue that these approaches rely greatly on deep learning algorithms and therefore suffer from certain weaknesses described previously.

3. Methods

3.1. The ANNABELL system

ANNABELL [32] is a cognitive architecture entirely based on a large-scale neural network (2M neurons). The system learns to communicate through natural language, developing language processing skills at the sentence level, starting from a supposed *tabula rasa* condition, i.e. no *a priori* knowledge on the structure of phrases or meaning of words [32,38]. The model provided a link between neural models of language and cognitive models of working memory [32]. The global organisation of the model complies with the multicomponent working memory (M-WM) framework [47], particularly on the role of the executive functions in language processing tasks, supported by evidence from experimental psychology, neuropsychology, and cognitive neuroscience [47].

Figure 1: **Schematic diagram of ANNABELL**. The system comprises a Long-Term Memory (LTM), a Short-Term Memory (STM), a Central Executive (CE) and a Reward System. All components are neural networks.

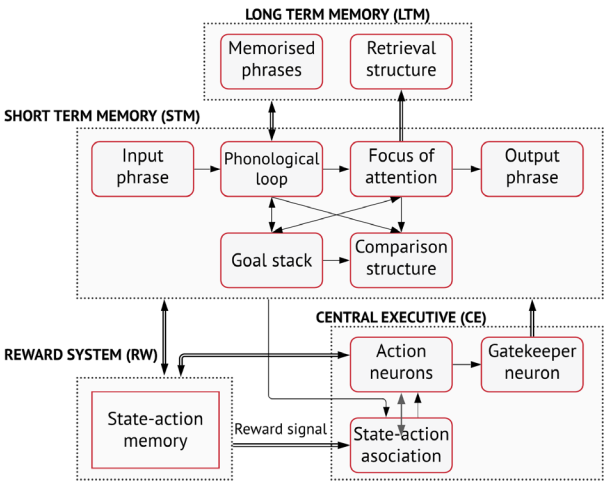


Figure 1 presents the four main components comprised in the model: a verbal short-term memory (STM), a verbal long-term memory (LTM), a central executive (CE) and a reward structure. ANNABELL is a pertinent implementation in AI of the concept of a central executive exploited for language learning, which allows to disambiguate and generalise on several different tasks and, unlike classical state-of-the-art DL models, to learn complex rules from only a few examples. The STM includes a phonological store, a focus of attention, a goal stack, and a comparison structure. The phonological store maintains the working phrase, which is either acquired from verbal input or retrieved from LTM. The focus of attention can hold up to four words. Goal chunks that contribute to decision-making are stored in the goal stack. The comparison structure recognises similarities between words in the phonological store, the focus of attention and goal stack, to aid the decision-making processes and support the generalisation capacities of the system. The working phrases are memorised in the LTM and are extracted by a retrieval structure using the focus of attention as a cue.

The CE operates the decision-dependent processes. These are not rule-based, but statistical decision processes. The CE receives in input a signal from the STM components (internal state) and outputs mental actions that direct the flow of information among the slave systems, through adaptive neural gating mechanisms. The statistical nature of the CE grants the generalisation property of the system. Intuitively, the CE comprises a state-action association system (SAA), a set of action neurons and gatekeeper neurons. The system processes verbal information using *mental actions*, which are elementary operations on word groups, phrases and other subnetworks, triggered by action neurons. The mental actions are represented by the flow of signal among different WM buffers, controlled by neural gates i.e. the gatekeeper neurons. The gating mechanism is regulated by the SAA. This is a neural network that takes as input a representation of the internal state of the working memory and sends output signals that open/close the neural gates.

Through the reward mechanism, used to train the SAA to associate mental actions to the internal state of the system, the model learns from examples how to control the flow of information among the slave systems and therefore, how to perform proper mental actions. This is a key feature of the model, given that the connections that are affected by the reward mechanism are connected to action neurons, rather than being directly connected to output words or phrases. Therefore, the system learns preferentially to build the output using sequences of elementary operations, in sentence level. This type of architecture underpins the generalisation capabilities of the system (see also Appendix 3).

3.2. Cognitive Learning Theory

Anderson et al. [48] proposed a hierarchy of knowledge consisting of declarative (factual) knowledge in the form of true affirmations, procedural (imperative) knowledge, which is the skill of performing some task and, metacognitive knowledge that describes the ability to use or relate past experiences in similar unseen tasks. This is applied in ANNABELL for training the system the skill of usage-driven language acquisition. The declarative knowledge of the system consists of a set of declarative statements and cues on using them, as naturally occur in the language. By training the model how to answer to simple questions, using simple example phrases via communicative interactions with the interlocutor, we build the procedural knowledge of the model

in language-related tasks. It complies with the Natural Approach of Language Learning [49], according which grammar rules are not essential when first acquiring the language. Rather, the continuous exposure to the language and how it is properly wielded in everyday situations, resembling a child's daily interaction with the parent, leads to spontaneous emergence of speech [49]. Similarly, though not intended to process speech, the methodology applied in this work, to train our system the acquisition of languages, does not focus on learning the grammatical constructions of the languages; instead, syntactical and semantical soundness are yielded by experiencing examples of how language is used in contextual verbal exchanges with the interlocutor. The generalisation skill emerges by following the same line of reasoning in sentence production and recalling these experiences in close endeavours (metacognitive knowledge).

3.3. The Dataset

In this work, the dataset is devoted to the thematic group *People*, originally described in [30]. It is in part inspired by the Language Development Survey (LDS) work of Rescorla et al. [39,40], which provides a valuable insight on the number and types of words or word combinations known and used spontaneously by toddlers. We use the LDS to construct a systematic dataset. Our focus is language acquisition and no classical tasks studied in NLP, thus we prefer to draw our corpus using vocabulary that occurs naturally in the process of language development, over standard NLP datasets. The generated dataset used here is suitable for an extensive quantitative evaluation, to model the acquisition of multiple languages in a cognitive architecture, using a simple vocabulary at developmental level of complexity.

The declarative sentences (factual knowledge) are used to describe the situated social environment of a fictional 4-year-old girl called Annabell, which includes twenty people and nine possible relationships. The human-system talk is modelled as question-answering. Our training methodology is inspired by parent-child verbal interactions of how parents use simple communicative examples to query the child about the world, rather than teaching the grammar constructions or specific linguistic forms. However, we do not model a real child's talk that emerges from real-world interactions. The answers given by the system assess its ability to process the acquired information and reasoning skills. The questions used in the dataset are also inspired by the work of Rescorla [39,40]. The English dataset is translated in the new languages, while not changing its content, amending sentences to address different morphosyntax or introducing where needed distinct uses of plurals, genders, noun cases, verb conjugations, "pro-drop" forms etc.

The human teaches the system to answer to question Q, by guiding it to build a valid answer (the state-action association that produced the answer is rewarded and permanently memorised). Some declarative (how-to) sentences give prescriptions on how to perform a task, e.g. to tell if someone is younger/older than you, you have to compare your age with theirs or the possessive pronoun for a girl is her. These are not sentences produced by the system (Table 1A, Appendix 1), but they aid sentence production. They comprise the knowledge stored in the LTM, which from a physiological perspective, is acquired from past experience (e.g. naturally using pronouns to refer to people and their possessions, without actual awareness on what a pronoun is). 46 tasks (questions) are used to train the system the skills of:

- using pronouns to refer to people and objects;
- answering polar, multiple-choice and wh-questions, e.g. *do you have a brother, what is your sister's name*;
- age comparison tasks, counting and comparing numbers: *who is older Tom or Susan*;
- telling its own likes and dislikes: *do you like to <verb>*;
- recognising other people's likes and dislikes (and the types of possible relations between the system and the persons): *does your father like to drive*;
- recognising different professions of different persons: *Dad is a teacher; the teacher teaches in the school*.

The content and organisation of the dataset People is same for each language. However, to endow the multi-linguistic competences of the system and compare its behaviour against the monolingual validation [30,36], it must be trained accordingly with the increased complexity of each language. For example, let us consider an answer that expects the use of the pronoun *"which"*. While in English *"which"* is gender-neutral, the other languages distinguish between at least two gender forms. Therefore, the teacher must provide more possible associations that will teach the system to use the pronoun correctly in different scenarios.

The dataset is divided in sets of training and test. The training procedure is organised in two stages. In the first stage, the system receives and memorises in the LTM a set of declarative sentences (explicit knowledge), in the form of verbal descriptions (e.g. *you have a mother*, *you are Annabell*, *you like dolls* etc.). In the subsequent training stage, the teacher asks a set of questions related to the declarative sentences and suggests a valid answer to each question, by guiding the system through a series of mental actions to build that answer (implicit knowledge). At least one of the interrogative sentences of each group is used for training, unless different languages require more examples of the same question type to generalise properly. In the test, the human uses the remaining interrogative sentences of each group, to assess the generalisation capability, i.e. to recall the information given by the declarative sentences to answer to questions close in structure and context to those trained, that however require using different nouns, pronouns, adjectives or verbs. We set a strong criterion for measuring accuracy, considering an answer valid only if it is syntactically and semantically correct and appropriate for the conversation.

Table 1 The number of declarative sentences used to set/describe the social environment, the number of interrogative sentences used for training and the number of questions in the test stage, for each language dataset.

Dataset ***People***	Declarative sentences	Interrogative sentences (train)	Interrogative sentences (test)
English	308	89	292
Greek	319	89	292
Italian	355	89	292
Albanian	364	89	292

3.4. Selected languages and their properties

The languages considered in this work are of different complexities, generally defined as the number and diversity

of elements, along with the intricacy of their inter-relational structure [41] and here regarded in the following aspects:

1. Syntagmatic/lexical complexity: word length, composite words structure, prepositions, different degrees of deixis.
2. Morphological complexity (word formation), e.g. full irregular plural nouns.
3. Organisational complexity; component arrangement (e.g. adjectival order) and the word order in sentences.
4. Semantic complexity devoted to the meaning of the words in the way they are arranged in the sentences.

Although the languages belong to (distinct branches of) the Indo-European family, a different language of those selected would only require a properly amended training corpus of phrases, with no change in the overall learning mechanisms. We do not focus on the language complexity or organisation, but rather how it affects the performance of the system during language acquisition. The linguistic competences of the system depend on how well it is trained to acquire each language. Building proper knowledge by experience, via ample input examples, requires identifying the linguistic expectations that underlie all language-related tasks. Some languages involve genders and verb conjugations that vary with the number of tenses and persons.

The Greek language features three gender types (masculine, feminine, neuter) and four cases (nominative, genitive, accusative, vocative) for nouns, while adjectives and articles agree in gender, number and case with their respective nouns. In most cases, the gender of the noun cannot be deduced by a rule, but it must be learned. The language can flexibly form compounds and tends to be periphrastic (usually for future tense) [43], affecting the generalisation capacities of the system as the meaning can depend on the different number of words or word order in a phrase. However, this monolectic compound-constructing capability of the Greek language (single compound words convey the meaning expressed by a sentence or paragraph) is often advantageous.

The Italian language shares similar features in terms of gender, number and case. Personal pronouns are not essential to the meaning and are often omitted, as the verb form itself indicates the subject (ho fame » I am hungry), unless when necessary for clarity or to add desirable emphasis or contrast [44]. Direct and indirect object pronouns (that receive direct or indirect actions), cannot stand alone without a verb [44].

In Albanian, articles are vital to the language as they combine with the noun to indicate the reference and specify the definiteness of the noun (usually four types). Adjectives are often accompanied by the connective article and vary in gender and number with the noun [45]. There are 6 noun cases (nominative, accusative, genitive, dative, ablative and vocative), introducing a change in the word structure that may be difficult to address. Plural is generally irregular.

These peculiarities suggest that each language imposes its own requirements in the training stage, in the structure of the dataset, the number and types of inputs needed for proper learning and how the system is trained to perform a task. Moreover, proper disambiguation is instrumental when they appear in coexistence during simultaneous acquisition. The sentences of the corpus are appropriate for conversation with a pre-school child, capable but no expert in delivering information in conventional ways, therefore simplifications are made. Grammar rules are preserved, however, when possible, the sentences are arranged uniformly in length and word order without syntactical violation. Most of the issues described here have been addressed in the dataset. That meaning, the system is exposed to ample quantities of comprehensible input required for each language that results in acquisition of the language (e.g. irregular plurals, genders).

4. Experimental Validation

The system is indifferent to the languages and does not have cross-linguistic awareness (translation competences). It performs the same mental actions at the sentence level (phrase memorisation, word extraction, etc.), regardless of the received language, i.e. it handles the sentences in the same manner. Hence, the architecture and information elaboration procedure remain intact, while languages can vary. However, proper acquisition of each language required training the system accordingly to the use of nouns, verbs, pronouns, etc., as naturally occur in each specific language.

For a fair comparison, the language datasets are uniformly organised to include equal training and test sets i.e. equal exposure. The number of declarative sentences that describe the social environment can be different for each dataset, as specific languages expect differences in the semantic meaning, for example when conjugations need be considered. We aimed to maintain same amount of training examples per task, to avoid having more learning samples in one language with respect to another (Table 1). Some languages required a finer training, to render the capability to recognise and generalise on their properties. E.g. in English, adjectives are neither gender nor number distinct, whereas the other languages feature both. Often a plural adjective used to describe a group of two or more people, with different genders, is generally masculine. If the system is asked the question "who *is older Tom or Susan"* and let *Susan be older*, upon receiving the adjective *older* expressed in masculine (Greek: megalyteros), the system should answer using the adjective *older*, in feminine form (Greek: megalyteri). Should the contrary occur and let *Tom be older*, the system must use the other gender form. The answer would not be both syntactically and semantically correct unless the system is trained in each scenario. Although this is not essential in English, both examples must also be learned, to ensure that none of the languages is over trained with respect to another, conversely the performance comparison would be flawed. Similar amends are made for all datasets. The structure of the datasets with these modifications is given in Table 2.

Table 2: The number of total sentences, total new words and average words per sentence, in each of the datasets, used to train and test the system. New words infer the different words in the whole corpus for each of the languages.

Dataset	Declarative sentences	New words	Words/sentence
English	308	213	6.049
Greek	319	240	6.611
Italian	355	227	6.265
Albanian	364	226	6.126

4.1. Cross-validation (CV) on individual languages

For a quantitative performance evaluation, we use a k-fold cross-validation technique. The CV procedure is organised

in four sessions (rounds), for each language dataset. The rounds are executed separately, starting from a clean state and the final performance of the system is then averaged on all four and for each language dataset. Each round follows three stages of execution: (a) the system first acquires the declarative sentences that set its experience in a text-based environment, (b) in the 2nd training stage, it learns to answer to different types of questions (arranged in 46 groups) and (c) its behaviour is then tested through a previously unused set of questions for each group, similar in structure as those learned during the 2nd stage of training. The accuracy is measured at the end of each round, before executing the next round anew. To obtain the four rounds independently, at each round, we extract randomly one (or more, when required by the specific language) interrogative sentence of each group, to build the training set (stage 2). We use the remaining others of the same group for the test, with the constraint that the same training question should not be used in different rounds. This allows assessing the system behaviour and performance four times independently, by varying the training and test sets. The order in which the tasks are learned and tested are randomised, e.g. task 46 can be learned before task 1 and test queries can appear randomly from each group.

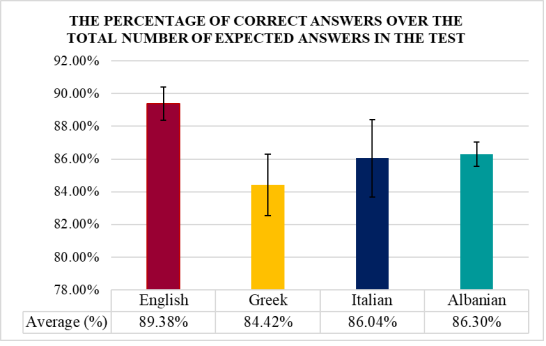
An important feature of the model, attributed to its architecture, is the ability to execute iterative mental operations during decision-making to build a valid answer. When asked to solve a task, i.e. answer to a question, the system needs to retrieve, extract and compare the phrase(s) from the LTM, that are most appropriate for the type of input query. The model does not exploit correlations among words or next word predictions, because they are represented by orthogonal vectors. Conversely, an answer is sometimes the result of following a line of reasoning that combines information from several declarative sentences. Let there be question Q of type *"is your <relationship> older than you"*.

Q: Is your friend older than you?
A: Yes she is.

To answer to question Q, it uses the learned past experiences:

1. Which person corresponds to the relationship (*Letizia is your friend*);
2. To count and compare numbers;
3. Its own age (*you are 4 years old*);
4. The age of the person referred by the relationship (e.g. *Letizia is five years old*);
5. The use of personal pronouns to address people (e.g. *"She"* instead of *"Letizia")*

Figure 2: Percentage of correct answers and the standard deviation, in each language dataset. The values are averaged over the four rounds of the CV.



In the test stage, the system should be able to generalise in similar questions to Q, on a new *<relationship>*, using all the past experiences it has acquired with respect to that *<relationship>* (e.g. who the person is, their age, etc.).

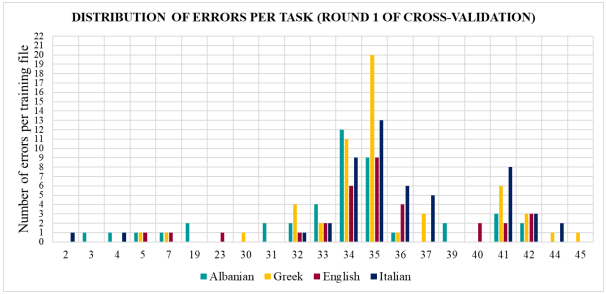
Figure 2 reports the average accuracies with their standard deviations: 89.38% in English, 84.41% in Greek, 86.04% in Italian and 86.3% in the Albanian language. The accuracy is defined in terms of ratio of the correct answers over the total number of requested output sentences, and is averaged over the four rounds of the CV, in each language dataset (see also Table 3). A correct answer is strictly both syntactically and semantically valid and appropriate for the conversation.

Table 3: The number of correct answers over the total number of requested sentences for all four rounds of the CV, in each language dataset. The answer is considered valid when it is both syntactically and semantically correct.

Dataset	Round 1	Round 2	Round 3	Round 4
English	260/292	264/292	264/292	256/292
Greek	237/292	254/292	247/292	248/292
Italian	240/292	259/292	259/292	244/292
Albanian	250/292	250/292	252/292	256/292

In Figure 3 are illustrated the distributions of errors for various language-related tasks, on one round of the CV. The errors for all languages have a uniform distribution across the test set, with peak values in specific tasks. Error-free tasks are not reported in Figure 3 and usually occur in the case of simple evenly structured sentences, of same length and word order, that conform to exact same grammatical rules (word/plural formation, noun clauses, …).

Figure 3: Distribution of errors for several tasks, in the first round of the CV. For all languages in all rounds, the errors have a uniform distribution across the test set, with steady peak values in tasks 34, 35, 41 and 42 and random peaks elsewhere. Error-free tasks are not presented here; they usually consist of simple evenly structured sentences, same in length and word order, applying the exact same grammatical rules.



Peak errors occur in tasks of type, *<person> likes to <verb>…* (action) and *<person> likes <noun>(s)* (object). During the 2nd training stage, the system is taught to answer to the question *"does <person> like …"*, in two cases: when they do and when they do not. The system ought to recognise the subject and the object/action related to the verb "like".

Let there be the following case: in the training stage, the system knows that "Mum likes to watch the TV".

Teacher: Does Mum like to watch the TV?
Annabell: Yes she does

The system is tested on a similar question, for which the correct answer is "no she does not". Instead, it answers:

Teacher: Does Mum like to drive the car?
Annabell: Yes she does

Because the words are not close in meaning (TV, car, etc), the system does not fully understand what is expected to do

i.e. recognise likes/dislikes. Instead, it responds similarly to what has learned in the training stage using a new object/action by virtue of its ability to form an action and recognise objects in its environment. To generalise on these tasks, the system should recognise the concept of "like/does not like" and transfer it to other situations that expect similar contextual meaning. Other tasks relate to the question "*what does <person> do*" and expect an answer of type *"<person> is a profession*". While the meaning can be perceivable for a human, it is rather challenging for the system to generalise on two sentences that do not share the same structure.

Although the errors are evenly distributed across the test set, typical errors are often related to some features of the language. A property of the Greek language is that masculine nouns (and male names) change their ending depending on the case, forcing in addition the article to vary or disappear altogether in the singular form according to case and gender; e.g. άντρα (man) would change with the case as follows: nominative: ο άντρας, accusative: τον άντρα, vocative: άντρα [43]. Questions that feature a noun case and expect an answer in another noun case, require a finer training process.

Other errors are related to the use of articles and irregular plural formation. A question is triggered in plural form and an answer is expected in singular form. With the plural being irregular, the system often cannot extract the root word (singular) of the noun to build a valid answer. The error fluctuations for the same task across the four rounds are often dependant on the learning example used in stage 2. The tasks in the dataset *People* are plausible for a simple intelligible conversation; however, a larger learning set is required to train a neural network the skill of proper acquisition.

4.2. Cumulative cross-validation

So far, we have demonstrated the ability of the system to learn independently and satisfactorily different natural languages, with distinct grammars and syntactic features, in different levels of complexity, coming at no cost of architectural changes and of the general neural processes that underpin how the system learns. Therefore, the system can be regarded as four monolingual subsystems.

The aim of this work is to evaluate if the system can acquire any language of choice and generalise properly on numerous tasks in multiple languages, when exposed to all languages jointly. That is, to what extent can the system separate and become aware of the (dis)similarities of the languages? How can it decide to respond in a multilingual environment, where it has acquired information expressed in different languages and forms and, whose expected behaviour must reflect correctly the type of conversation and the language it is being delivered? Does learning languages jointly affects the ability to acquire a language satisfactorily and comparably to its monolingual state? We are inspired by the literature on multilingualism [46] and the human cognitive capability to process multiple languages simultaneously.

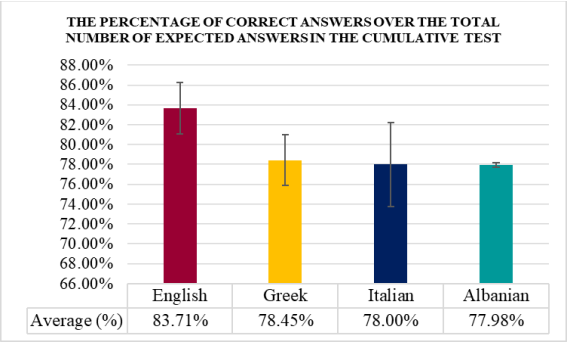
We performed a four-round open-ended cumulative cross validation. This relatively long developmental process is sustained by the large-scale of the network and the ability of the system to perform real time communication. The rounds are built similarly as explained in section 4.1, however here each round includes all four language datasets (e.g. the individual 1st rounds of English, Greek, Italian and Albanian now comprise the 1st round of the cuulative CV and so on). Thus, in each round, the system was trained consecutively on all languages, before querying it with random language at random order. This was repeated independently for each round and the final performance was averaged over the four rounds. In a single round of the cumulative CV the system received a total of 1346 declarative sentences (stage 1), was trained with 356 interrogative sentences (stage 2) and tested with 1168 previously unused questions, expressed in all four languages. The compound dataset contained 900 different new words, with an average of 6.263 words per sentence, with respect to all languages.

Table 4: The number of correct answers over the total number of requested output sentences in each language dataset, for four rounds of the cumulative cross-validation training.

Dataset	Round 1	Round 2	Round 3	Round 4
English	231/292	250/292	253/292	243/292
Greek	222/292	243/292	224/292	227/292
Italian	227/292	218/292	249/292	217/292
Albanian	227/292	229/292	227/292	228/292

There was a slight decrease in the overall accuracy compared to the individual training. In Figure 4 are shown the total accuracies averaged over the 4 rounds, was 83.81% for English, 78.51% for Greek, 78.00% for Italian and 77.97% for the Albanian language (see also Table 4). While the three latter unveil lower accuracies, the overall decrease is comparable for all languages (Greek 5.91%, English 6.05%, Italian 8.04% and Albanian language 8.33%). Languages other than English share similar results devoted to their complexity, as explained in section 2.3. The results indicate the competence of the system to separate the languages at the grammatical and semantic level, and deliver appropriate conversation in the expected language, with a success rate of above 77% in each of the languages.

Figure 4: Percentage of correct answers over the total number of expected answers, & the standard deviation, evaluated on each language datasets. The values are averaged over the four rounds of the cumulative cross validation.



We expected the fall, as the sizes of the compound datasets were four times larger than the monolingual datasets, thus affecting the number of learnable interconnections that are created and loaded from the system. The links saturate much faster, which may cause a mild interference in the processing of the acquired information that can be overwhelmed by increasing the number of input examples in the training stage and/or increasing the number of neurons and connections.

In some cases, the answers entailed mixing languages, with little or no change in the conveyed meaning of the output for

the type of conversation. We observed three cases: (a) the response does not answer the question, but is syntactically correct, expressed in a mixed language (it uses an equivalent pronoun in another language); (b) the answer is syntactically & semantically correct and appropriate for the conversation, but expressed in a different language from the query; (c) the answer is semantically sensible yet not syntactically correct; e.g. the system mixes two equivalent words of two different languages, within the same phrase (App. 2). Though "code-switching" is regarded in literature as a natural part of (multi)bilingualism [49] there is no strong evidence that this is the case in ANNABELL, hence we considered the answers erroneous. However, it is particularly interesting to note that there exists some sort of relevance on the above scenarios, at either the grammatical (syntactic or morphological) level or the word (semantic) level. It might indicate that the system understands the context of the task or some correlation between the languages and switches or "borrows" words to construct a sound sentence. Further work is required to conclude appropriately on the above.

4.2.1. Incremental training & testing

To investigate deeper if the multilingual system can manage to learn the languages to the same degree of skill as learning each of them alone i.e. at monolingual skill, we performed an incremental cumulative language training. Given that the results of cumulative CV can be affected considerably by the size of the corpus as compared to the number of neurons in the subcomponents and the number of connections among them, training the system incrementally grants a greater understanding on the linguistic competences of the system. We grouped the languages in four datasets: monolingual, bilingual, trilingual and quadrilingual, each sequential dataset comprising the languages of the previous lower-level dataset. For each dataset independently, we performed a cumulative training, before the test.

Table 4: Incremental language training. Accuracy is measured in terms of correct answers over the total requested sentences, for each language, in the 1st round of CV, when training is performed with 2 languages (Albanian, Italian), 3 languages (Albanian, Italian, Greek) and all 4 languages.

Accuracy/lang.	*1 lang.*	*2 lang.*	*3 lang.*	*4 lang.*
Albanian	85.62%	80.48%	80.48%	77.74%
Italian		80.14%	80.82%	77.74%
Greek			78.42%	76.03%
English				79.45%

The results are given in Table 4, on the 1st round of the CV. As the number of sequential combinations with two, three and four languages is large, the proposed scenario, here assumed significant, first comprised the languages of higher complexity with lower accuracies in the cumulative CV: Albanian (mono), Albanian & Italian (bilingual), Albanian, Italian & Greek (trilingual), all languages (quadrilingual).

The results of incremental training showed rather higher performance of the system, with little or no variance in accuracy, when two and three languages are learned. The system can generalise properly using articles, nouns, verbs, adjectives, and other open-class words, as naturally expected in each language. This suggests that acquiring a language jointly with others does not significantly affect the ability to acquire the language as skilfully as acquiring it alone. As discussed earlier, the mild drop from an additional fourth language may be related on one side to the interference among the information acquired in different languages and, on the other side, on how the limitations in neuron and connection numbers affect storing and processing very large datasets.

5. Narrative Story Comprehension on Preschool Literature

Machine comprehension of texts and the ability to answer context questions is an open problem in AI and the human-machine interaction. From a psychological perspective, understanding narratives requires assessing what people recall from the story and their response to probe words [51]. In this session, we describe an experiment with the model, aiming to explore its potentiality to capture the meaning of a short narrative story. This is not a traditional machine reading comprehension of any text e.g. from public datasets or benchmarks, but instead is limited to the level of child comprehension. The text is taken from the book "My first jungle story" [52], of preschool literature, translated in our four languages of choice, to construct the necessary datasets.

The subject involves 11 animals; the main character Leo the lion undertakes a trip to meet other animals, to which it asks a lot of open-ended questions. Unlike earlier datasets in ANNABELL, the declarative sentences used to describe each of the animals communicate a continuous coherent and cohesive meaning in the story. The system is trained to answer to three consecutive questions on each animal. The phrases are originally extracted from the book: (i) what did Leo learn about the <animal>, (ii) can it <verb> and (iii) what can it do. The questions are logically linked and convey little meaning when asked alone. Although text comprehension also presumes making prepositions that derive from the story and previous existing knowledge in the LTM, which is often not found within the text itself [51], this is not a target of the experiment. Rather, we focus on exploring three main aspects of text assessing:

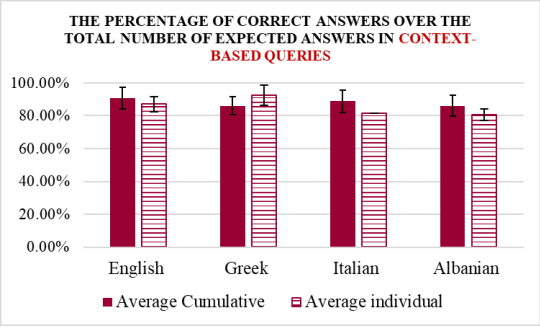
1. The competences of the system to give appropriate meaningful answers within a certain context.
2. The ability to dialogue with the human and, logically and chronologically link questions, while tracking the referent. Evaluate if the system understands which animal the teacher refers to when (a) addresses a question using "it" instead of the animal name, but suggests an action as clue (*can it <verb>*), or (b) asks a general question without reference on the animal (*what can it do*), i.e. any animal is an answer candidate; the system ought to find which fits the given context, i.e. recall past queries. Finally, describe animals using full-length sentences.
3. The generalisation capabilities of the system on a similar but previously unseen test set on the story, in all the languages, acquired both separately and cumulatively.

The system learns the story through a set of declarative sentences (stage 1) and to answer questions about one animal (stage 2), whilst it is tested on the remaining others (App. 2). The "what" - questions have more than one candidate answer, i.e. any of the animals in the story, however the system can choose the referent that relates to the past questions. In the second question, the system never "sees" the animal and the

action (described by a verb) together in a sentence but can recall the referent from the first question and link it to the action given in the second query. Moreover, this question can be formulated specifically for the animal (and can be of any length), without affecting the ability to generalise. Whereas the third question is a tricky one. How does the system know which is the referent animal? How can it relate this question to the prior two? The test results show that the system is able to track the correct referent across queries, use meaningful descriptions and generalise appropriately on each character. The system can use the pronoun "it" and animal name interchangeably. The property to dialogue in such manner is attributed to its ability to store a sequence in the goal stack (typical in cognitive architectures) when a mental action cannot be performed immediately. During decision making in later stages, the system can recognise that one word in the phonological store is equal to a word of the phrase stored in the goal stack and use this link to build a valid answer.

The 4-rounds CV is performed as explained in sections 4.1 and 4.2. The training sets for each round are built using a new animal for learning while the 9 remaining animals are used in the test. The rounds are executed independently. Both methods, individual and cumulative, perform roughly the same, apart from few cases of code-switching observed in the latter. Due to identical spellings in different languages (e.g. Hippo: *Ippopotamo* in Italian & Greek), the model misplaces the pronoun "it" but answers correctly at semantic and syntactic level (App. 2).

Figure 5. Cumulative and individual training on text comprehension. First, we evaluated the monolingual behaviour, where the system learns the story and answers to related questions in only one language, for each language apart. In the cumulative approach, the system learns the script in all the languages and answers to questions in different languages randomly, in the test stage. The results show the percentage of correct answers, out of 27 total expected answer i.e. 3 x 9 animals, and the standard deviations. The system can answer to a question, even when not all three are answered correctly.



Even when the system fails to answer to a precursor query, the rest of the dialogue might not be affected e.g. the system cannot build an answer to the first question, but identifies the goal task and uses it to handle correctly the conversation that follows (App. 2). The system tracks the referent in different parts of the conversation.

The results in Figure 5 show that both approaches give satisfactory performances above 80%, with the standard deviation comparable in both cases. The are some statistical fluctuations which can derive from the small number of samples in the training and testing stage. We would expect the system's accuracy to be higher in monolingual training, which occurs only in the case of the Greek language (errors in cumulative training result from the code-mixing of some common spellings explained above). Training the tasks cumulatively strengthens its ability to generalise.

6. Conclusion

In this work, we assessed the potential of the developmental cognitive architecture, ANNABELL, to acquire and dialogue in multiple natural languages simultaneously. The study was performed using four languages (Greek, Italian, Albanian, English) with peculiar structure and complexity.

The system was priory trained and tested in each language alone, performing in cross-validation at 84-89% for different languages. This was compared with a 4-rounds cumulative CV, in which, the system was trained jointly on all languages, before testing, yielding accuracies between 78-84% across the different languages. The aim was to verify if the model could learn the languages together to the same degree of skill as learning each apart. The results show robust generalisation capabilities of the model. In the cumulative CV, its answers matched the language of the query, despite having acquired and stored the information of all languages jointly. The methodology was inspired by the literature on the natural organisation of information in the brain of early bilinguals who store the acquired information in the same area of the brain without specific distinction of the language [57].

Via a deeper investigation on the progressive cumulative learning of one or more languages, through an incremental training approach, the system only experienced a slight decrease in the performances as more languages were added in the training stage. The results showed that there were no significant accuracy variations when learning two or three languages, but outcomes were lower in the fourth language. Tests suggest that this might derive from a slight interference in the processing of the information given in different languages, which can be overwhelmed by increasing the number of input examples in the training stage and/or increasing the number of neurons and connections; however, a conclusive discussion is a target of future work. This will attempt to verify if the performance depends on the number of learnable connections, for a predefined architecture, on the sequential order in which languages are learned and, what is the confusion that can occur in multiple language acquisition.

Finally, we explored the competences of the system in multilingual context questions, on a short story of preschool literature, to assess its ability to elaborate narratives and dialogue with the human coherently & cohesively in linked questions. The individual and cumulative approaches performed above 80% in all languages, proving its capacity to track the referent in different parts of the conversation.

In-depth investigation of how varying the training and the focus of the attention, on different tasks, could be useful to improve the model's performance. In the intended upgrade of the system, we will study how the nonverbal signals during language acquisition can be associated with the focus of attention of the architecture and their role in learning. In future works, ANNABELL embodied in a humanoid robot will experience language grounding and account nonverbal signals available from the human interlocutor during natural communication, extending to appropriate datasets from observed speech transcripts and other HRI corpora.

7. References

[1] D. Silver, et al. Mastering the game of go with deep neural networks and tree search. Nature, 529(7587), 2016, 484–489.

[2] D. Amodei et al., Deep speech 2: E2E speech recognition in english & mandarin; Intern. Conf. on Machine Learning, 2016, 173–182.

[3] H. Fang et al. From captions to visual concepts and back; in: IEEE conference on computer vision & pattern recognition, 2015, 1473-1482.

[4] A. Karpathy et. al Deep visual-semantic alignments for generating image descriptions, in: Proceedings of IEEE conference on computer vision and pattern recognition, 2015, 3128-3137.

[5] Z.Gan et al. Semantic compositional networks for visual captioning, in: IEEE Conf. on computer vision & pattern recognition, 2017, 5630-5639.

[6] Q. Wu et al. Image captioning and visual question answering based on attributes and external knowledge. IEEE transactions on pattern analysis and machine intelligence, 40(6), 2018, 1367-1381.

[7] Z. Yang et al, Stacked attention networks for image question answering, in: IEEE Conf. on computer vision & pattern recognition, 2016, 21-29.

[8] H. Xu et al, Enhancing semantic image retrieval with limited labelled examples via deep learning. KBS journal, 163, 2019, 252-266.

[9] B. Mitra, et al, Learning to match using local and distributed representations of text for web search, in: Proceedings of the 26th International Conference on World Wide Web, 2017, 1291-1299.

[10] X. Fu et al., Semi-supervised Aspect-level Sentiment Classification Model based on Variational Autoencoder. KBS Journal, 2019.

[11] F. Huang et al, Image-text sentiment analysis via deep multimodal attentive fusion. KBS journal, 2019.

[12] F. Zhang et al., Collaborative knowledge base embedding for recommender systems, in Proc. of the 22nd international conference on knowledge discovery & data mining, 2016, 353-362.

[13] G. Wang, et al, HAR-SI: A novel hybrid article recommendation approach integrating with social information in scientific social network. KBS, 148, 2018, 85-99.

[14] E. Dupoux, Cognitive science in the era of AI: A roadmap for reverse-engineering the infant language-learner. Cognition, 173, 2018, 43-59.

[15] Google AI Blog, Teaching Google Assistant to be Multilingual. 2018. https://ai.googleblog.com/2018/08/Multilingual-Google-Assistant.html

[16] IBM Cognitive Insight: IBM Watson is now fluent in nine different languages (and counting), in Connecting the cognitive world. 2016. http://www.wired.co.uk/article/connecting-the-cognitiveworld.

[17] B. M. Lake et al. Human-level concept learning through probabilistic program induction. Science, 350(6266), 2015, 1332-1338.

[18] B. M. Lake et al., Building Machines That Learn Think Like People. Behav Brain Sci, 2016, 1-101.

[19] R. Navigli, Natural Language Understanding: Instructions for (Present & Future) in Proceed. of the 27th Intern Conf. on AI, 2018, 5697-5702.

[20] M. A. Kelly et al, How Language Processing can Shape a Common Model of Cognition. Procedia compsci, 145, 2018, 724-729.

[21] Derek M, James A. R, Neural architectures for learning to answer questions, Biologically Inspired Cognitive Architectures, Vol 2, 2012, pp 37-53, ISSN 2212-683X, https://doi.org/10.1016/j.bica.2012.06.002

[22] R. Miikkulainen, L. Elman, Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory, MIT Press, '93.

[23] R. Miikkulainen, Script-based inference and memory retrieval in subsymbolic story processing. Applied Intelligence 5(2), '95, 137–163.

[24] P. Fidelman, et al A subsymbolic model of complex story understanding, in: Proceedings of the Cognitive Science Society, 7(27), 2015, 660-665.

[25] P.F.Dominey, Recurrent temporal networks and language acquisition—from corticostriatal neurophysiology to reservoir computing. Frontiers in psychology, 4, 2013, 500.

[26] X. Hinaut et al Real-Time Parallel Processing of Grammatical Structure in the Fronto-Striatal System: A Recurrent Network Simulation Study Using Reservoir Computing. PLoS ONE, 8(2), 2013, e52946.

[27] X. Hinaut et al. Cortico-Striatal Response Selection in Sentence Production: Insights from neural network simulation with Reservoir Computing. Brain and Language, vol. 150, Nov. 2015, pp. 54–68.

[28] X. Hinaut et al. A Recurrent Neural Network for Multiple Language Acquisition: Starting with English & French. Cognitive Computation: Integrating Neural & Symbolic Approaches, CoCo @ NIPS 2015

[29] X. Hinaut et al., Teach Your Robot Your Language! Trainable Neural Parser for Modeling Human Sentence Processing: Examples for 15 Languages. IEEE Transactions on Cogn. and Developmental Systems, vol. 12, no. 2, pp. 179-188, 2020, doi: 10.1109/TCDS.2019.2957006.

[30] J. Gaspers et al, Constructing a Language From Scratch: Combining Bottom–Up and Top–Down Learning Processes in a Computational Model of Language Acquisition. IEEE Transactions on Cognitive and Developmental Systems, vol. 9, no. 2, 183-196, 2017, doi: 10.1109/TCDS.2016.2614958.

[31] Clément Moulin-Frier, Tobias Fischer et al., DAC-h3: A Proactive Robot Cognitive Architecture to Acquire and Express Knowledge About the World and the Self, IEEE Transactions on Cognitive and Developmental Systems, 2017, DOI: 10.1109/TCDS.2017.2754143

[32] B. Golosio, et al, A cognitive neural architecture able to learn and communicate through natural language, PloS ONE, 10(11), 2015.

[33] Xiaowei Zhao et al. Bilingual lexical interactions in an unsupervised neural network model, International Journal of Bilingual Education and Bilingualism, 13:5, 505-524, DOI: 10.1080/13670050.2010.488284

[34] Matusevych, Y et al (2017). The impact of first and second language exposure on learning second language constructions. Bilingualism: Language and Cognition, 128-149.

[35] Spandana Gella et al. Image Pivoting for Learning Multilingual Multimodal Representations, 2017 Conference on Empirical Methods in NLP (EMNLP), pg 2839–2845 Copenhagen, Denmark.

[36] Ákos Kádár et al. Lessons Learned in Multilingual Grounded Language Learning, in the Proceedings of the 22nd Conference on Computational Natural Language Learning, pp 402–412 Brussels, Belgium, 2018

[37] Josje Verhagen et al. How do verbal short-term memory and working memory relate to the acquisition of vocabulary and grammar? A comparison between first and second language learners, Journal of Experimental Child Psychology, Volume 141, 2016, Pages 65-82

[38] Giorgi I., Golosio B., Cangelosi A., Masala G., Annabell a Cognitive System Able to Learn Different Languages, New Trends in Intelligent Software Methodologies, Tools and Techniques, 2018, 992 – 1003.

[39] Rescorla L, Alley A (2001) Validation of the language development survey (LDS): a parent report tool for identifying language delay in toddlers, J Speech Lang Hear Res 44(2): 434–445.

[40] Rescorla L, Achenbach T (2002) Use of the Language Development Survey (LDS) in a National probability sample of children 18 to 35 months old, J Speech Lang Hear Res 45(4): 733–743.

[41] B. Szmrecsanyi, B. Kortmann, Introduction: Linguistic complexity: Second Language Acquisition, Indigenization, Contact, October 2012,

[42] Golosio B.,et al: A cognitive neural model of executive functions in natural language processing, *in (ICBICA 2015),* vol. 71 (196-201), 2015.

[43] M. Poulopoulou, Modern Greek: Grammar Notes for Absolute Beginners - A User-Friendly Grammar for Levels A1-A2, Uni. of Crete.

[44] S. Peyronel et al, Basic Italian: A Grammar and Workbook, Taylor & Francis e-Library, 2005, Master e-book ISBN 0–415–34717–3

[45] Victor A. Friedman, Albanian Grammar, Studies on Albanian and Other Balkan Language Peja: Dukagjini. 2004.

[46] De Houwer, Annick, Two or More Languages in Early Childhood: Some General Points and Practical Recommendations. ERIC Digest, ERIC Clearinghouse on Languages and Linguistics Washington, 1999.

[47] Baddeley AD: Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology* 63: 1-29, 2012.

[48] Anderson et al.. A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objective. New York: Longman Publishing, 2001.

[49] Krashen, S. D, T. D. Terrell. The Natural Approach: Language Acquisition in the Classroom. California: Alemany Press, 1983.

[50] N. F. Ramirez, Why the baby brain can learn two languages at the same time, The Conversation, University of Washington, 2016.

[51] Robert S. Wyer, Jr (ed) *Knowledge and Memory: The Real Story*. Hillsdale, NJ. Lawrence Erlbaum Associates. 1-85, 1995.

[52] Brown Watson, My first jungle story, ISBN 9780709725961, 2019

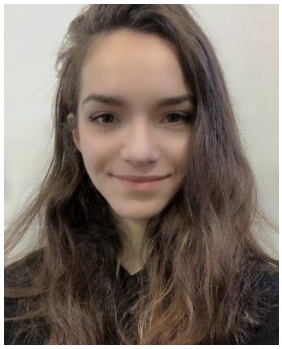
[53] Jefferies Franchise Note, Creating Shareholder Value with AI? Not so Elementary, My Dear Watson, July 12, 2019.

[54] Oberauer K Access to information in working memory: exploring the focus of attention. Journal of Experimental Psychology. Learning, Memory, and Cognition 28(3): 411–421, 2002.

[55] Bryck RL, Mayr U On the role of verbalization during task set selection: switching or serial order control? Mem. Cognit. 33(4), 611–623, 2005.

[56] Vandierendonck A. Role of Working Memory in Task Switching. Psychologica Belgica, 52(2–3), 229–253, 2012.

[57] Abutalebi J, et al. The Bilingual Brain as Revealed by Functional Neuroimaging. Bilingualism: Language & Cognition 179, 2001



IOANNA GIORGI is a postgraduate research student at the Department of Computer Science and part of the Cognitive Robotics Lab Research Group (COROLAB) at the University of Manchester (UK). Her research interests are in multiple natural language understanding and grounding for developmental robotics, artificial social companions and, human-robot and robot-robot cooperation. She is working on the framework of cognitive neural models for language development and their applications in robotics.
ORCID: 0000-0001-9583-6959

BRUNO GOLOSIO is associate professor of Applied Physics at the University of Cagliari, Italy, since 2016, and coordinator of the School of Medical Physics of the University of Cagliari since 2017. His main research interests are computational neuroscience, artificial neural networks, machine learning and computational methods for biomedical imaging, high-performance computing. Over the last two decades, he participated in several research projects, with increasing level of responsibility. Currently, he is principal investigator of the project icei-hbp-2020-0007 in the Interactive Computing E-Infrastructure for the Human Brain Project (ICEI-HBP). He collaborates on the Wavescales (wave scaling experiments & simulations) experiment in the framework of the Human Brain Project and on the AIM (Artificial Intelligence in Medicine) project founded by the INFN (Italian Institute for Nuclear Physics). He is the first author of the cognitive neural model of language development, ANNABELL and the software library NeuronGPU for fast simulation of large-scale networks of spiking neurons. He contributed to numerous software for computational methods in biomedical physics. ORCID: 0000-0001-5144-6932

MASSIMO ESPOSITO is currently a researcher at the Institute for High Performance Computing and Networking of the National Research Council of Italy (ICAR-CNR). He is a contract professor of Informatics at the University of Naples "Federica II", Faculty of Engineering, since 2012. He is responsible for the "Cognitive Systems" laboratory at ICAR-CNR since 2016. His current research interests are in the field of Artificial Intelligence (AI), with focus on AI algorithms and techniques, combining deep-learning and knowledge-based technologies, to build intelligent systems able to converse, understand natural language and answer to queries, with emphasis on the distributional neural representation of words and sentences.
He is involved in numerous national and European projects. He is member of the program committee of several international conferences and workshops and, currently member of the editorial board of various international journals. He is author of over 100 peer-reviewed publications on international journals and conference proceedings.
ORCID: 0000-0002-7196-7994

ANGELO CANGELOSI is Professor of Machine Learning and Robotics at the University of Manchester (UK). He also is Turing Fellow at the Alan Turing Institute London, Visiting Professor at Hohai University and at Universita' Cattolica Milan, and Visiting Distinguished Fellow at AIST-AIRC Tokyo. His research interests are in developmental robotics, language grounding, human robot-interaction and trust, and robot companions for health and social care.
Prof. Cangelosi has produced more than 300 scientific publications, had led many UK and international projects (e.g. THRIVE, EnTRUST, APRIL, BABEL, ROBOTDOC, ITALK) and has been general/bridging chair of numerous workshops and conferences including the IEEE ICDL-EpiRob Conferences. Cangelosi is Editor of the journals *Interaction Studies* and *IET Cognitive Computation and Systems*, and in 2015 was Editor-in-Chief of *IEEE Transactions on Autonomous Development*. His latest book "Cognitive Robotics" (MIT Press), coedited with Minoru Asada, will be published in 2021. ORCID: 0000-0002-4709-2243

GIOVANNI MASALA is Senior Lecturer in Computer Science and Leader of the Robotics Lab in Manchester Metropolitan University (UK). He is also Visiting Research Fellow at University of Plymouth (UK). His research interests are in Artificial Intelligence (AI) and Robotics, natural language understanding, social robots for elderly and AI in medical applications.
Dr. Masala has produced more than 80 scientific publications in international journals and conference proceedings. He is involved with numerous international research grants and is leading the U.K. partnership of the EU Project Interreg 2 Seas Mers Zeeën (2014-2020) "AGE Independently" (AGE'IN). He has been part of program committees and has chaired several international workshops and conferences. Masala is Guest Associate Editor of the journals *Frontiers in Computational Intelligence in Robotics*, *Frontiers in Computer Vision* and in *Applied Sciences SI*.
ORCID: 0000-0001-6734-9