Human Language as a Tool for Conceptual Development in Cognitive Robotics

A DISSERTATION SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE AND ENGINEERING

2022

Ioanna Giorgi

Department of Computer Science
Submitted Thesis: *Human Language as a tool for Conceptual Development in Cognitive Robotics*

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>5</td>
</tr>
<tr>
<td>List of Figures</td>
<td>6</td>
</tr>
<tr>
<td>Abstract</td>
<td>10</td>
</tr>
<tr>
<td>Declaration</td>
<td>11</td>
</tr>
<tr>
<td>Copyright</td>
<td>12</td>
</tr>
<tr>
<td>Acknowledgement</td>
<td>13</td>
</tr>
<tr>
<td><strong>Chapter 1</strong></td>
<td>15</td>
</tr>
<tr>
<td>Preface</td>
<td>15</td>
</tr>
<tr>
<td>1.1. Context</td>
<td>15</td>
</tr>
<tr>
<td>1.2. Aims &amp; Objectives</td>
<td>16</td>
</tr>
<tr>
<td>1.3. Research Hypotheses</td>
<td>17</td>
</tr>
<tr>
<td>1.4. Contribution to Knowledge</td>
<td>17</td>
</tr>
<tr>
<td>1.5. Research Outputs</td>
<td>18</td>
</tr>
<tr>
<td>1.6. Thesis Structure</td>
<td>19</td>
</tr>
<tr>
<td>1.6.1. How to read this thesis</td>
<td>20</td>
</tr>
<tr>
<td><strong>Chapter 2</strong></td>
<td>22</td>
</tr>
<tr>
<td>Literature Review</td>
<td>22</td>
</tr>
<tr>
<td>2.1 Robotic Paradigms</td>
<td>22</td>
</tr>
<tr>
<td>2.1.1 Cognitive Robotics</td>
<td>22</td>
</tr>
<tr>
<td>2.1.1.1 Vygotskyan Cognitive Robotics</td>
<td>23</td>
</tr>
<tr>
<td>2.1.2 Developmental Robotics</td>
<td>24</td>
</tr>
<tr>
<td>2.2 Language and Development</td>
<td>25</td>
</tr>
<tr>
<td>2.2.1 Language in Cognitive Development</td>
<td>25</td>
</tr>
<tr>
<td>2.2.2 Computational and Robotic Language Models</td>
<td>27</td>
</tr>
<tr>
<td>2.2.2.1 Models of high-level cognition</td>
<td>27</td>
</tr>
<tr>
<td>2.2.2.2 Robots that use language</td>
<td>28</td>
</tr>
<tr>
<td>2.2.2.3 Other language processing domains</td>
<td>30</td>
</tr>
<tr>
<td>2.2.3 Multilingual Computational Models</td>
<td>32</td>
</tr>
<tr>
<td>2.3 What is Conceptual Development</td>
<td>33</td>
</tr>
<tr>
<td>2.3.1 Concepts and their Understanding</td>
<td>33</td>
</tr>
<tr>
<td>2.3.2 The Path to Conceptual Development</td>
<td>35</td>
</tr>
<tr>
<td>2.3.3 Embodied Conceptual Development</td>
<td>36</td>
</tr>
<tr>
<td>2.3.4 Working Memory and Conceptual Development</td>
<td>37</td>
</tr>
<tr>
<td>2.3.4.1 A Working Memory Compliant Framework</td>
<td>38</td>
</tr>
<tr>
<td><strong>Chapter 3</strong></td>
<td>41</td>
</tr>
<tr>
<td>Modelling Multiple Language Learning</td>
<td>41</td>
</tr>
<tr>
<td>3.1 Context</td>
<td>41</td>
</tr>
<tr>
<td>3.2 Language and Development</td>
<td>2</td>
</tr>
</tbody>
</table>
Chapter 3: Conceptual Development in Cognitive Robotics

3.2 Prologue
3.2.1 The Hierarchy of Knowledge Representation
3.2.2 The Selected Human Languages
3.3 Multilingual Modelling Methodology: Part 1
3.3.1 Dataset 1: A fictional multilingual social environment
3.3.2 Experimental Validation
3.3.2.1 Monolingual Acquisition
3.3.2.2 Polylngual Acquisition
3.3.2.3 Piecemeal Cumulative Learning
3.3.3 Multilingual Modelling Methodology: Part 2
3.3.4 Dataset 2: A multilingual narrative passage
3.3.5 Experimental Validation
3.3.5.1 Monolingual Acquisition
3.3.5.2 Polylngual Acquisition
3.3.5.3 Piecemeal Cumulative Learning
3.4 Human-Robot Bilingual Interaction: Part 3
3.4.1 Robotic artefact
3.4.2 The Instantiated Robotic Model
3.4.2.1 Automatic Language Recognition and Speech Synthesis
3.4.2.2 Yet Another Robotic Platform (YARP)
3.4.2.3 The Flow of Information in the Cognitive Robotic Model
3.4.3 The Behavioural Experiments
3.5 Conclusion

Chapter 4: From Human Language to Robot Behaviour
4.1 Context
4.2 Prologue
4.2.1 Workaround the Visuo-Spatial Sketchpad
4.2.2 The Competence for Runtime Learning
4.2.3 ON-World Embodiment
4.3 The Learning Methodology
4.3.1 Pre- and Post-Linguistic Primitives
4.3.2 Binding Domain-Specific Information
4.3.2.1 From Concrete to Abstract Concepts: Learning Snowballing
4.3.2.2 On-Line Open-Ended Runtime Learning
4.4 Robot Learning Experiments
4.4.1 Perception Tasks
4.4.1.1 Target Training Data
4.4.1.2 Runtime Object Learning
4.4.1.3 Target Test Data
4.4.2 Manipulation Tasks
4.4.2.1 Target Training Data
4.4.2.2 Runtime Action Learning
List of Tables

Table 1 The multilingual corpus. The number of declarative sentences used to set/describe the social environment (column 2), the number of interrogative sentences used for training (column 3) and the number of questions used in the test stage (column 4), in each language (column 1). .................................................................45

Table 2 The fraction of valid answers produced in each language in each round of the cross-validation. A valid answer must be both syntactically and semantically sound. .........................................................................................................................47

Table 3 The fraction of valid answers produced in each language in each round of the cumulative cross-validation. A valid answer must be both syntactically and semantically sound. ..............................................................................................................51

Table 4 The corpus of phrases used to train and test the robotic system the capacity to map language to perception. The test lexicon was generated at runtime using printed images of animals and speech descriptions. 80

Table 5 The verbal explicit knowledge maintained in the long-term memory (LTM) of the cognitive architecture. It comprises information about objects that belong to the category of containers (exemplars) and of objects that do not (non-exemplars). Non-exemplar objects possess close-in attributes with the containers, e.g., the hat-trick has an opening is similar to the jug has an opening (column 2). Moreover, the robot knows a range of liquids, which are discriminable based on their colour attribute. ..............................................................................................................87

Table 6 The knowledge representation in long-term memory (LTM). The robot has memorised the cross-lingual pivot representations from preliminary training (column 3). These associations of different semantics co-exist in memory and the explicit knowledge from different languages is intertwined. The robot self-retrieves the relevant lexical term using the language indicator (_en, _it). Notice how there are no direct representations of central concepts used here in the Italian language (e.g., bottle, jug,… or colours: transparent, red…). The event constraints (the actor’s beverage preferences) are made known to the robot in the form of declarative phrases in the Italian language. Auxiliary phrases on the food preferences are introduced to assess with certainty the robot’s generalisation capacities during decision-making. *The translation of the Italian phrases is only shown to improve the reader’s comprehension but are not part of the information stored in the LTM of the learning framework. ..............................................................................................................96

Table 7 The one-occasion learning sample and generalisation sets used in the foundation task. Test case 1 includes cases of category members with similar criterial attributes, which meet the event constraints. Test case 2 includes cases in which the colour constraint of the liquid is not met. The robot was able to execute the task successfully for all test case 1 and cease the task otherwise (test case 2). .................................................................................................99

Table 8 The one-occasion learning sample and generalisation sets used in scenario 1. Test case 1 includes examples in which the involved concepts possess attributes that meet the event-imposed constraints. Test case 2 includes cases of non-exemplars. The robot successfully executed the task in all test cases 1 or ceased the task otherwise (case 2). ........................................................................................................................................................................101

Table 9 The one-occasion learning sample and generalisation sets used in scenario 2, In English and Italian. Test case 1 refers to examples of concepts that meet the constraints in the context. Test 2 involves cases in which these constraints are not met. In both linguistic scenes, the robot was able to solve all examples of test case 1 and cease the task appropriately in test case 2, by means of self-inference on the critical constraints. ..............................................................................................................103
List of Figures

Figure 1 The theoretical Working Memory models of Baddeley & Hitch (up left) and Cowan (up right). The schematic diagram of the ANNABELL system architecture (adapted from Golosio et al., 2015) (down). ...............................39
Figure 2 The three major types of knowledge as defined by Anderson & Krathwohl. (2001). ...........................................42
Figure 3 The original (English) and polylingual corpora. The language datasets are normalised to involve equal lexicons and learning samples. ........................................................................................................................................46
Figure 4 The quotient of valid answers produced by the learning framework in each language. The accuracy values are averaged over the four rounds of the cross-validation and the dispersion is indicated by the standard deviation. 47
Figure 5 Distribution of errors across question types in each disjoint language corpora and all four rounds of the CV. All languages have a uniform error distribution spanned on all rounds, with steady peak values in complex tasks 34, 35, 41, and 42 and random peaks elsewhere. Error-free tasks are not shown; they recurrently relate to phrases of simple structure, of equal length and word order, and unvarying grammar. .............................................................................................................49
Figure 6 Sample illustration of grammar & morphosyntax dissimilarities between English (up) and Italian (down). ..........................................................................................................................................................50
Figure 7 The quotient of valid answers produced in each language when trained jointly. The values are averaged over the four rounds of the cumulative CV and the dispersion is indicated by the standard deviation. The contrast with the monolingual acquisition is displayed to assess at what degree of skill (in terms of accuracy) does the learning framework acquire the languages jointly over each alone .................................................................52
Figure 8 The elaboration competence of the model in augmenting multilingual lexicons. The proposed scenario assumed as significant first accounted languages of higher intricacy, gradually scaling up the dataset with more languages, one at a time. The end performance of the quadrilingual dataset was faithful to the results of round 1 in the cumulative cross-validation (column 1, Table 3), suggesting that the chronology in which languages are learned is imperceptible .................................................................................................................................................................53
Figure 9 Context-oriented mono and polylingual acquisition of a narrative story. The ability to learn the story in a mono language was compared against learning it parallelly in multiple languages. The accuracy is averaged over the four rounds of cross-validation and its fluctuation is indicated by the standard deviation. In the multilingual scenario, the learning framework was probed randomly in any language, attending to the criterion that related questions must be of the same language, whereas the language can vary with the queried animal ..................................................................................................................55
Figure 10 The Q.bo One robot, (The Corpora, 2022) ..........................................................................................................................................................57
Figure 11 Bridging the cognitive architecture with the Q.bo One robot via YARP communication protocol and Google Cloud Platform. The flow of information to and from the cognitive architecture and the robot is shown. On the left side, the robot manages speech recognition, automatic language detection and speech synthesis, based on the predicted language code from the user’s speech. It feeds information to YARP over SSH, which runs altogether with the cognitive architecture on a remote computer (right side), responsible for the information processing at a neural level. ...............................................................................................................................................58
Figure 12 Illustrative samples from the live experiments with the Q.bo One robot, handling bilingual dialogue with the human in English and Italian simultaneously. ......................................................................................................................................................59
Figure 13 High-level flowchart of the learning methodology to acquire new and adapt to the changes in information occurring at runtime from real-world interactions. The blocks of the diagram represent the neural components of the cognitive model: WM-Working Memory (STM in ANNABELL), LTM – Long-Term Memory, CE – Central Executive. ..................................................................................................................................................64
Figure 14 The ON-World embodiment of the cognitive model ANNABELL with the NAO robot ..................66
Figure 15 Binding domain-specific information. A) Sensory stimuli from the non-linguistic domain (visual, motor) are conventionally represented through a linguistic symbol. They are uniquely mapped to their respective linguistic forms (linguistic domain). ⟨_form form⟩ mappings are memorised permanently in the long-term memory (LTM). B) Meaning can be retrieved from form and vice versa. ...........................................................................................................68
Figure 16 The proposed learning methodology to integrate language with perception and action. The fundamental neural components of the learning framework are shown: the STM (Short-Term Memory), which handles multimodal input phrases and output production; the CE (Central Executive), responsible for the statistical decision-dependent
processes on neural level processing (instead of a conventional IF-Else logic); the LTM (Long-Term Memory), which stores multimodal information that scales up during the runtime interaction. The CE can disambiguate between the task being learned or queried, along with the type of query requested in input. Should the instruction implicate concrete concepts, the robot self-maps the linguistic concept with the internal representation of the concept (visual or motor) and retains the association in long-term memory. Should the instruction involve abstract concepts, the robot uses linguistic experience and input from the user, to map the concept indirectly to earlier-learned (concrete) concepts that have a representation in memory. Learning of novel concepts (objects, actions) occurs at runtime, which involves learning feedback for cases of incomplete scenes. If a concept is queried before learning it, the robot hints the human without breaking the learning process. The human can teach the concept using verbal descriptions and live captured images or motor exploration, similar to child learning.  

Figure 17. The snowballing artifice. Action primitives are programmed in the robot’s motor map as predetermined trajectories (grasp, lift, drop). The human guides the robot through simple natural language instructions to combine the motor primitives suitably to produce a new outcome. While elementary actions are directly hooked to a specific motor exploration, complex actions expressed by more abstract verbs (take, place, make, add) entail convoluted motor modalities and their meaning is indirectly mapped to a certain combination of lower-order actions.  

Figure 18 Object Learning. The robot is shown an object (ball) and verbally dictated its linguistic term (this is a ball). The classification label from the vision module and the recognised speech are jointly employed in the learning framework, which self-generates the association <ball ball> that is permanently retained in long-term memory (LTM). The newly acquired mapping can be uniquely and directly retrieved from memory in future endeavours.  

Figure 19 Action Learning. The human triggers a motor exploration on the robot (e.g., grasp) and dictates its meaning (e.g., this means to grasp). The robot retrieves the action trajectory and generates the respective pseudo-sensory representation, which is combined with the recognised speech and processed jointly in the learning framework. The WM of the framework self-generates the association <grasp grasp>, which is permanently retained in long-term memory to be readily accessible for future use.  

Figure 20 Manipulation verbs that relate to primitive (concrete) actions are directly mapped to their internal sensory representation and the associations (<_action action_>) are permanently maintained in long-term memory. Human instructions that involve such actions (“step X do something”) are autonomously elaborated by the robot, which converts them to low-level symbolic sequences (representations of concrete objects and actions). Human instructions that involve abstract actions that do not have a direct representation in the workspace and, hence, not an <_action action_> association in memory, are extrapolated by logical linking of their lower-level constituents (indicated in steps). Two low-level action primitives are implicated in a 1st order of a high-level sequence. Higher-order manipulation verbs are generated by adding and combining duly more steps, in an incremental toolchain (2nd high-level sequences, 3rd high-level sequences, …).  

Figure 21 Handling sensory imperfections during the runtime collection of target test data. In Perception Tasks (A), the framework uses a combination of user speech and a vision stimulus recorded by NAO’s cameras and classified by AlexNet. To obtain ground-truth sample inputs at the learning framework, the human tutor lightly filters potential sensory uncertainties. The human re-captures the input in the event of a speech (content words) or vision misrecognition. The accuracy is only tested at the ends of the learning framework (over the end-to-end robotic system). Similarly, in Manipulation Tasks (B) the framework receives a pre-filtered compound input of speech samples and trajectory information from the robot's motor map. The robot's motor information is collected and self-mapped to its symbolic representation without human supervision. The motor stimulus is critical for the construction of the test toolchain from low-to-high level instructions.  

Figure 22 The accuracy rate for perception tasks. The tested vocabulary is not anticipated at design time but acquired preferentially during runtime interactions with the human. The human probed the robot with similar phrases to those trained involving the animal names dictated during the interaction. Without re-training the framework on the scale-up corpus, the robot could generalise with an overall accuracy of 94%, demonstrating a robust competence for symbolic mapping, open-ended learning of novel instances not specified in the initial pre-training, and reasoning on assorted interactive tasks.  

Figure 23 The tested vocabulary is acquired gradually at runtime. The robot learned novel objects, actions and mappings, which were retained in memory in a readily accessible state for task-solving. It could construct grounded sequences for high-level (abstract) actions by transferring low-level constituents across multiple level domains and integrating language with the physical body. Language facilitated the learning of abstract concepts, which were indirectly self-mapped to their meaning by linguistic experience and their relation to earlier-learned concepts of lower levels. Starting from a simple motor map, the robot refined its motor modalities by creating new action representations aided by language in close liaison with the human and the workspace and learning to solve complex tasks by doing.
The robot would converge the meaning of (abstract) instructions elicited from the current scene (available tools) and the outcome of the interaction (make – a tea, a toast, or a juice). .................................................................82

**Figure 24** Demonstration of the NAO robot preparing tea with a mug, a bottle, and a teabag. .................................................83

**Figure 25** The one-occasion learning and the runtime test snowballing toolchain to convert natural language (NL) instructions into robot actions. Lower-order primitives are self-combined duly by the cognitive architecture, by following step-by-step NL explanations from the human tutor. At every increasing order of abstraction (darker colour tones), the robot recalls relevant information from long-term memory and verifies the constraints of the event. Should these constraints not be met, or information cannot be retrieved from the memory, the robot ceases the task. In the test, <object> can be any of the alternative containers (cup, jug), or random non-container objects with close non-critical attributes; <liquid> can be any of the liquid exemplars and <colour> any of the liquid colours (Table 5). ...89

**Figure 26** Discerning examples from non-examples of [container](s). The robot is trained with a single example to take the container it sees in front of it. The spoken instruction “take the container” is combined with the percept data received by the virtual robot in the Webots world. The learning framework that elaborates the input (linguistic + non-linguistic), uses its knowledge from long-term memory to validate if the perceived object possesses the attribute of container. In the test, the robot must discern the objects appropriately, based on this property. Non-examples of containers share close-in attributes with the containers; however, these attributes are non-critical in the context. The robot is tasked with retrieving only relevant information (critical attribute) that responds to the human’s request. ...90

**Figure 27** Training of the concept-learning task. The robot is trained with a single example to add water to the cup from a bottle. For visualisation, the blue colour in the figure represents water (transparent). The speech data (add water to the cup) is combined with the percept data from the observed scene (utensil: bottle, colour: transparent, target object: cup). The latter are represented conventionally through linguistic symbols. The robot verifies two constraints: a) is the utensil a container, b) is the liquid observed in the scene water (by colour assessment). Should these constraints be met, the robot uses the snowballing toolchain to retrieve the earlier-learned primitive actions to complete the task. Otherwise, it ceases the task. In the test, the task is repeated consistently by replacing the bottle with other containers (jug, cup) holding liquids of random colour. .................................................................91

**Figure 28** Scenario 1: Concept attainment of increasing abstraction. The robot is trained with a single example to prepare tea using a bottle, given the verbally dictated lexicalised attribute “to make tea you need water” (LTM). The previous experience of adding water to the cup is not re-trained but retrieved directly from past experience with the concepts from the foundation concept-learning task, which are memorised permanently in long-term memory. In the test, scenario 1 is repeated consistently, using other utensils (jug, cup) with different liquids. The robot must colour-discriminate liquids and execute the task only with exemplars of water containers. .................................................................92

**Figure 29** Cross-lingual pivoting. A concept (percept or motor data) is used as a common-ground pivot to the different lexical terms that are used to label the concept in different languages. The <meaning, form> pairs in each language correspond to different (multimodal) knowledge sources. The transposing between a source to the other is achieved through the pivoting concepts with the help of the language-specific indexes (_en, _it). The different semantic labels are only indirectly linked among one another, through the common-ground symbolic representation of the concept they refer to (i.e., no translation). .................................................................94

**Figure 30** Scenario 2: Conceptual content in polyglot contextualisation. The robot is trained with one example (one per language) to serve to some person a drink of their preference. The scenario unfolds from the concept-learning task, with the novel lexicalised constraint of discerning liquids based on their colour AND the lexicalised attribute of “preferred drink”. That means previous experiences trained and memorised in the concept-learning task are retrieved directly to execute the task upon verifying the requisite conditions. In the test, the robot-model is consistently tasked with serving a drink to a certain actor: which drink is self-determined by the robot based on the lexicalised constraint of preference and its identity in the scene is self-verified via colour observation. The robot demonstrates identical behaviour in each of the languages, albeit having very limited experience in Italian (no snowballing), using concepts as common ground to retrieve the learned experiences. .................................................................97

**Figure 31** Foundation Task: Concept learning by doing. The objects in group A belong to the category of containers. Group B includes objects that possess close-in non-critical attributes with the containers, but are not containers (e.g., “the hat-trick has an opening” resembles “the jug has an opening”). This knowledge is stored in long-term memory. The first event constraint is to select only examples of containers (group A) and reject all non-examples (group B). The second constraint is to separate liquids based on their discriminable colour. An example of the concept refers to the queried liquid (speech) matching its observed colour identity in the scene (group D); otherwise, it refers to non-examples (group E). .................................................................98
Figure 32 Scenario 1: Concept attainment of increasing abstraction. The robot is trained to make tea using a bottle of water. Learning to use the involved tools (cup, teabag) is achieved via the snowballing artifice. Group A includes instances that can be used to prepare tea, whereas group B involves instances that cannot. The linguistic input alone can successfully narrow down the available concepts into those that fit the running context, by abstracting from their differences and similarities (C1, C2) in new encounters not anticipated at learning (C3).

Figure 33 Scenario 2: Conceptual content in polyglot contextualisation. The robot is tasked with serving a preferred potable to a person (up left). The preferences are formulated in simple Italian phrases (upright). Further food preferences are dictated to assess the ability of the robot to discriminate the concept of [drink]. Group A includes examples of concept appropriateness, in which the event-imposed constraints are met. Group B includes cases in which the observed drink does not match the anticipated identity of the actor’s beverage preference. The task is repeated two times: (a) robot queried in English and (b) robot queried in Italian.

Supplementary Figure 1 Schematic diagram of the ANNABELL learning framework. Each rectangle represents a subnetwork, composed of interconnected artificial neurons. Only the main high-level subnetworks are shown in the diagram. The arrows that join the rectangles represent directional connections among neurons of different subnetworks and buffers. During the acquisition phase, the input words are mapped onto neurons of an input-word buffer and are combined into an input phrase through proper indexing. In the association phase, the input phrase is copied to a working-phrase buffer in the phonological loop and is mapped to a neuron of the memorised-phrase buffer. The phrase is stored in the long-term memory (LTM) by changing the weights of the connections from that neuron to the working-phrase buffer, through a discrete version of the Hebbian learning rule. During the exploration and exploitation phases, word groups are extracted from the working phrase buffer and used by the focus of attention as cues to retrieve memorised phrases from the LTM. When the working phrase represents an action in the mental action sequence that cannot be performed immediately, it is copied to the goal stack, where it is stored until the action is carried out. The comparison structure is a subnetwork that recognises the presence of similar (equal) words in the working phrase, the phonological loop, and the goal stack. The elementary operations (mental actions) on word groups, phrase buffers and other SSMs (sparse signals maps) are triggered by the action neurons. The action neurons control the state of the gatekeeper neurons, which, in turn, manage the flow of signals between the subnetworks by acting as gates that allow or inhibit the information to pass. When the exploration leads to the desired output, the human interlocutor can trigger a reward. In the reward phase, the system retrieves the action sequence that led to the reward and memorises the association between the state of the system in each step of the sequence and the corresponding action. The state-action association is memorised by changing the weights of the input and output connections of the state-action association subnetwork through the k-winner-takes-all rule.

Supplementary Figure 2 The synaptic gating mechanism in the ANNABELL learning framework to gate the information transmission between the action neurons and neuron sets of the other SSMs.
Abstract

We are nearing a prospect in which robots will be an imminent constituent of our society. Humans will be symbiotically empowered by AI (Artificial Intelligence), presuming that these smart artefacts are devised to mimic human mental and somatic attributes. Human-robot synergy can be nurtured by a solid foundation of mutual understanding and seamless communication. Alas, current efforts to devise apt cognitive robotic models focus on low-order cognitive phenomena alone (perception, manipulation, motor coordination, navigation). Drawing on a hypothesised developmental paradigm of human cognitive functions, they have seized prominent aspects of embodied and situated cognition, which roots in motor behaviour and the environment. Yet they offer no clear sight how their blueprints can explain or scale up to high-level cognitive competence. These efforts often overlook a major component that sits at the core of our (human) social interaction: language. Language may be the natural interface between humans and robots, not only as the naive way we communicate our thoughts with others but having the overarching benefit of supporting cognition and intelligence.

This dissertation takes direct inspiration from theoretical psychology and revolutionary cognitive robotics perspectives. It seeks to address the near absence of high-level cognitive modelling in the current standpoints of cognitive robotics, advocating to appraise human language as a versatile cognitive tool that can support and enhance cognition. This thesis describes a series of cohesive research efforts conducted in augmenting stages aimed at involving language in robot cognition. Stage 1 models language proficiency in a cognitively sound manner that is closer to how humans develop and/or elaborate language. The significance of this modelling is that it can be used to study other humanlike cognitive aspects, in rooting to brain-inspired principles over fabricated solutions that focus on high accuracy for ad-hoc tasks. This stage models not one but multiple languages jointly, as multilingual competence is assumed to promote cognitive, neural, and social benefits, and to have a far-reaching impact on cognitive control. Yet, language cannot be fully understood if not viewed in relation to our perception, actions and interactions with the environment and the organisms in it. Thus, stage 2 models the symbolic mapping between language, body, and the environment to study how words get their meaning, i.e., how they manifest in the real world as visual and somatosensory information. In the proposed modelling, the learning of words, actions or both journeys from concrete to abstract concepts, where abstractness is a language attribute. With increasing abstractness, the number, and types of actions constructed in response to language to manifest that abstractness increases in a continuum. The results of this novel learning artifice with a humanoid embodied robot suggested that language has an impact on action learning and adaptive behaviour. These assumptions are forged ahead in stage 3 to study the impact of language on further aspects of cognition, such as the ability to categorise, abstract and voluntary control behaviour. Challenging experiments with a humanoid robot demonstrated that language could influence such phenomena. The robot could fathom concepts expressed in distinct languages (cross-lingual cognitive control), showing that labels become part of conceptual representations and can trigger those representations just as well as perception and action. Language appeared to boost cognition by promoting higher-level abstract reasoning, which required properties that were inferred rather than directly observed.
Declaration

No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or another institute of learning.
Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example, graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocInfo.aspx?DocID=24420), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in The University’s policy on presentation of Theses.
Acknowledgement

The people and occurrences that have aligned in my doctoral voyage are worth more than grandeur words. I treasure the unconditional love and invaluable advocacy that I have received to reach this milestone.

I am deeply thankful to my family, my parents Luljeta and Vasil and my sister Katerina, whose selfless abiding sacrifice, wise counsel, and benevolent ear have shaped my personal and professional growth.

I wish to extend my deepest gratitude to my academic supervisors, prof. Angelo Cangelosi and Dr. Giovanni Masala for taking the leap of faith in me from day one and for continuing to challenge, encourage, and mentor me with professionalism and kindness. I acknowledge the indispensable contribution of Dr. Bruno Golosio and Dr. Massimo Esposito in the conduct of this research, for the expertise I have received and for the cordial collaborative environment. My special regards to the Department of Computer Science of the University of Manchester, in which I gained an invaluable experience of work attitude and ethics, to mention, Dr. Barry Cheetham, who I had the pleasure to assist as a graduate teaching assistant.

Lastly, I would like to recognise the significance of my PhD cohort, in particular the Cognitive Robotics Lab (CoRoLAB) who shared this meaningful journey with me both as colleagues and beloved friends.
To auntie’s little peanut, Zoe (Ζωή)
Chapter 1

Preface

1.1. Context

The human brain has a remarkable ability to use information in creative ways to build and pass on knowledge, which grants human organisms' plasticity to evolve in their ambience. The capacity of adapting to miscellaneous environments is devoted to our cognitive ability (Boyd et al., 2011). People do not possess all the necessary information in all bearings but exploit their accumulated life skills while being exposed to continuous novel stimuli, which allows them to develop, grow and adjust to more complex practices. The attempt to make robots an integral part of our dynamic social structures and interactions, naturally imposes the need that they achieve similar competence. A fundamental form of cognition that favours humans, and could potentially enable robots to adjust as efficiently, is grouping concepts pragmatically into coherent categories. Successful learning and utilisation of categories allow us to draw non-trivial inferences on situations in which we lack direct experience and, as such, propelling understanding (Bruner & Austin, 1986). Solid theories on child’s psychology and behaviour advocate that, concepts and the act of categorisation, deeply reflect a child’s development, problem-solving and word-learning (Lupyan & Bergen, 2016). This conceptual development starts from perceptual clustering (e.g., prelinguistic infants grouping objects by colour) and progresses to nontrivial abstract thinking, which requires a fair amount of language (Markman, 1989; Sloutsky & Deng, 2019).

Social interaction and cooperative intelligence are necessary for contemporary AI (Dafoe et al., 2021). Two focal aspects of such intelligence include understanding and communication, which would enable machines to act in congruence with humans’ intentions and expectations. This requires a significant understanding of human language and other signals we use to communicate ideas. It is innate to want to investigate the critical role of language, not only as a communication system but its implication in the complex process of decision-making. Subjective forms of expression (“I prefer this”), allow many forms of reasoning that are arduous without linguistic concepts and can help grasp indirect intents. However, learning language alone cannot lead to understanding meaning, i.e., its real-world manifestation. The synergy between language and the tangible environment involves both proficiency of language structure and use (requisite 1) and mastery of grounding it in the world (requisite 2) (Bender & Koller, 2020). Cangelosi & Schlesinger (2018) claim that the physical body is instrumental to achieve this goal. Inspired by child psychology, they reason that an embodied agent should regulate the interactions among body, brain, and environment, to acquire somatic and mental skills in its physical surroundings. The robot should be able to map language appropriately to perceivable and tangible referents and extend the same capability across more abstract notions that are characterised by increased complexity and meaning variability with a weaker direct anchor in motor experience, and arguably stronger anchor onto language (Evans, 2006). Evidence advocates that there is a continuum between concrete and abstract concepts (Borghi et al., 2017). The more words become detached from tangible entities that are perceived through senses, the higher their abstractness (Wiemer-
Chapter 1: Preface

The continuum is a powerful tool to extrapolate strongly grounded experiences to more complex practices and thus, might help devise robots that can use concepts in a high-level way to represent knowledge (on experiences, objects, actions) similarly to humans (requisite 3). This knowledge can be materialised in words, creating structures that themselves represent instances of perception, manipulation, and progressive thought, as part of conceptual development (Clark, 2006).

When speaking about intelligence, either natural or artificial, it is nearly impossible not to end up alluding to language in some way. Thereupon, the question that governs the research described in this dissertation is: where does language edge in the humanlike cognitive development and the emerging vision of the intelligent embodied agent? Current efforts to build apt cognitive robotic models often neglect the influential role of language in manlike learning. Instead, it is in the belief of this thesis that language is a fundamental cognitive function that complements several forms of neural processing and guides context-sensitive thought and reasoning. On this premise, the offered research hypotheses are drawn from the incentive to investigate how language can spur embodied and situated cognitive modelling into higher-level forms of cognition, placing significant emphasis on conceptual content rather than the underlying representations of language.

1.2. Aims & Objectives

This dissertation aims to computationally model aspects of high-level cognition in robotics. To fulfil this aim and answer to the research question posed above: where does language edge in the humanlike cognitive development and the emerging vision of the intelligent embodied agent?, the research described here embarks on the efforts to account for human language in the effective learning of robots, considering language of similar (if not equal) significance as the perception and the physical body of the robot.

This thesis trusts that the successful symbiosis between language and the body for the emergence of conceptual development can be achieved by meeting the requisites 1-3 listed in section 1.1. As such, these requisites become the objectives of this research, as follows:

O.1. Computationally model language learning and proficiency (requisite 1).

i. Start with a comprehensive investigation of the existing literature of theoretical, scientific, and engineered models, to capture prominent aspects and potential limitations that would allow identifying the best policy to model language in a cognitively sound manner and approximate an artificial agent to the natural way humans learn, over engineered crafts.

ii. Model the competence to master language use in more than one human language (multilingual cognition).

iii. Demonstrate the potential of this modelling in extensive developmental computational experimentation.

O.2. Devise a symbolic capacity to associate internal mental representations of phenomena with linguistic labels (requisite 2).

i. Propose a methodology to ground language in the robot’s physical locale to fathom all the natural communicative signals that influence cognition: the auditory, the visual, the somaesthetic. This requires surveying current efforts in literature around language grounding and nourishing these to the cognitively plausible perspective adopted here.

ii. Demonstrate the potential of this symbolic mapping between language and the body in profound developmental experimentation.
O.3. Model high-level cognitive aspects in a robot, such as real-world learning, categorisation, abstraction, and autonomous control (requisite 3).

i. Investigate the theoretical insights on how language, in overt or covert form (inner speech), can empower cognition, and computational models that have managed to capture some of these phenomena in robotics.

ii. Propose a methodology to develop and structure the conceptual content of the robot, which mimics the development of psychological representations of knowledge in humans, i.e., how humans represent knowledge and use it in problem-solving, reasoning, and further learning.

iii. Demonstrate via proper developmental experimentation if a robot with sufficient conceptual development can regulate its behaviour in the workspace autonomously and intelligibly.

1.3. Research Hypotheses

The requisites/objectives raise key hypotheses on if and how an objective can be achieved in an AI model. Each research hypothesis is pitched sequentially from the intended objectives and governs a (methodology) chapter. A chapter/hypothesis may posit manifold implicit questions.

RH1: Can we devise multilingual cognition in a biologically sensible AI model congruent with how the bilingual brain is activated in response to language and how it resolves language conflict?

RH2: How can a cognitive model develop the symbolic capacity to associate internal mental representations of real-world phenomena (meaning) with their labels (language)? How do abstract words take their meaning in the tangible environment? Can a model learn novel praxes from the workspace constraints and in interaction with other organisms in it, which can be used directly for problem-solving without re-training the entire model de novo?

RH3: Can an artificial cognitive model develop appropriate conceptual content and structure its knowledge representation (i.e., organise its experience into coherent patterns) to draw non-trivial inferences on its locale? What is the cognitive role of language in this conceptual development? When can we assume that such models understand the concepts they use?

1.4. Contribution to Knowledge

This dissertation contributes to the existing literature in several respects through or reflected in the peer-reviewed publications (1.5). The impact of the research is reported by chapter.

1. It offers substantial contribution to modelling the acquisition of multiple natural languages using a computationally efficient neural model that mimics aspects of verbal information elaboration in the human brain. It distances itself from engineering solutions of natural language processing (NLP), instead being a precursor to promote further work on multiple interactive languages in specific domains. The research into understanding the underlying processes of multilingual cognition can be extended to vast potential domains beyond language development. The demonstrated endowment of the model in a robot can propel multilingual human-robot interaction, requiring only low-cost social companions. This first contribution (Chapter 3, publications iii, iv) foreruns the efforts to devise polyglot robots that can use languages as rich sources of information and transfer contexts to/from multiple tongues without loss of meaning (chapter 5).
2. It introduces enhanced learning mechanisms, novel knowledge representation and memory retrieval processes to handle task-oriented multimodal stimuli (verbal, visual, motor) in an on-line open-ended mode. The research presents the prime robotic instantiation of a complex language architecture drawn on the principles of the working memory. It proposes an original artifice to represent language of emergent abstractness into direct (concrete notions) and indirect (abstract notions) manifestation of perception and action in the physical workspace. The artifice enables the robotic model to learn new practices in/from the environment, using an unanticipated vocabulary that scales upon runtime from the interactions with the human. It can gradually amend its knowledge and self-representations of the workspace, using the newly acquired information directly to solve unfamiliar tasks with human guidance. This contribution (Chapter 4, publication ii) models the symbolic capacity to map language with internal mental representations that are needed to shape high-level cognition and launch a developmental learning snowballing blueprint leveraged in the (to come) concept-like behavioural tasks: the robot’s primitive dexterities extrapolate into complex motor control in close liaison with its locale and the human (child-like learning) (chapter 5).

3. It investigates the implication of language in the manifestation of high-level cognitive competencies, such as categorisation, abstraction, and voluntary control. It offers to the absent literature a learning model that adapts favourably to new occurrences in the real workspace, with no pre-determined logic, unattainable myriad corpora, or a significant amount of a priori training. Due to the emergence of conceptual content, the model can readily recognise new exemplars and utilise them directly to produce similar intentional outcomes as their categorial counterparts, as if they were the same kind of thing. Language further excels this ability, allowing many forms of abstract reasoning. This is explicitly observed through the original artifice proposed in this thesis to contextualise the robot’s behaviour in multiple human languages. Contextualisation allows validating whether the robot can analyse a new event in terms of the concepts surrounding it and draw inferences on unanticipated semantic bearings in which it lacks direct experience. This prompts the overarching benefit of understanding context-dependant tasks. This contribution is presented in publication (i) and detailed in Chapter 5.

1.5. Research Outputs

The research outputs that relate to the contributions of this thesis are listed below. The publications introduce the original research of the author of this thesis. The co-authors of these research outputs have contributed with dawning conceptualisation and qualitative supervision.

**Book Chapter:**


**Journals:**

1.6. Thesis Structure

This thesis is organised into 6 chapters.

Chapter 1: Preface

This chapter establishes the context of the research described in the dissertation, the posed research questions, and their significance. It draws on the key research, introducing its ambit and critical positioning with respect to existing proposals. By identifying the gaps and/or limitations of current research, it motivates why the conducted research makes a promising impactful contribution.

Chapter 2: Literature Review

This chapter reviews the body of literature that inspired and supports the research described in this thesis. The discussed bibliography is framed in a general context of cognitive psychology and cognitive sciences. It is divided into three parts:

Part 1: Robotic Paradigms
This part highlights an eminent body of literature on robotics research, narrowing the ample field of robotics in the two paradigms that fit the ambit of this thesis: Cognitive Robotics and Developmental Robotics.

Part 2: Language and Development
This section focuses on the role of language in human cognitive development. It discusses several ways by which language improves cognition, for learning, attention, memory, categorisation, and abstraction. The section reviews noted theories from cognitive psychology and computational models that have attempted to study these phenomena in artificial artefacts, in mono/polylingual

---

1 This output marks the point of departure of my doctoral studies and research vision.
domains. Furthermore, it discusses assorted efforts of computational modelling of language into perception and action, in line with the embodied perspective of cognition.

Part 3: What is Conceptual Development
This part discusses theories of human conceptual development. It articulates a thorough definition of concept and categorisation, and what it means to understand a concept according to learning sciences research. It introduces the notion of the working memory (WM), which is assumed to be the cauldron of concept formation in cognitive psychology.

Chapter 3: Modelling Multiple Language Learning
This chapter models the capability of a cognitive developmental model to jointly process and produce multiple human languages of sizeable intricacies. The chapter aims to contribute to the computational understanding of the characteristics that favour multi-language development, which is closer to how humans learn, inspired by findings on language control in the multilingual brain.

Chapter 4: From Human Language to Robot Behaviour
This chapter proposes a strategic acquisition of the meaning of linguistic stimuli of increasing level of abstractness, by anchoring them in the robot’s self-representation of the external world and body. The contribution of this chapter is two-fold: 1) it draws on the first robotic instantiation of a complex language architecture based on the principle of the Working Memory model; 2) it feeds open-ended learning, by which the robot acquires new behavioural repertoire continuously from task-oriented interactions in unanticipated scenes (i.e., past the training stage).

Chapter 5: High-level Cognitive Modelling in Robots
This chapter explores the role of language in modelling the high-level cognitive skills of robots. It proposes an original approach to conceptual development via contextualisation of robot behaviour in poly human languages. Contextualisation can validate the ability of the robot to analyse an event in terms of the concepts that surround it and draw inferences on unanticipated semantic encounters in which the robot lacks direct experience. It allows the robot to transpose the concept-like behaviours between cross-linguistic scenes with little training, where language stimuli drive the attention to the relevant aspects of the task and activate the mental states of the motor experiences.

Chapter 6: Final Thoughts and Prospect
This chapter recapitulates, synthesises, and concludes the research described in this dissertation. It reformulates and describes the scientific contribution and impact of the research. The chapter further discusses open problems and research questions for the near and far sight, with what regards modelling human-like intelligence and understanding of machines and robots.

1.6.1. How to read this thesis
This section helps the reader understand the logical link and coherence of the research described in the methodology chapters 3-5, in that they are part of an augmenting and chained approach to achieve the cardinal aim of the research: modelling high-level cognitive phenomena in cognitive robotics. The outcomes of each chapter support the assumptions and hypotheses for the next chapter. The overall research conducted in this stage-like methodology exploits solid theories from several disciplines (cognitive psychology, linguistics, artificial intelligence, neurorobotics) to introduce its own proposal of situating and contextualising poly human languages to model and
demonstrate high-level cognitive phenomena in robotic artefacts (poorly addressed or lacking altogether in the existing literature). The steps to achieve the aim follow a developmental robotics paradigm exerted on a brain-inspired cognitive architecture and can be summarised as follows:

1. **Chapter 2** introduces the reader to the central notions around robotics and psychological insights on human language and cognitive development that are required to understand the efforts of the research described in this dissertation.

2. **Chapter 3** presents the research that is drawn from the ambition to tackle objective 1 (O.1) and is conducted in response to the research hypothesis 1 (RH1).

3. **Chapter 4** reveals the efforts to meet objective 2 (O.2) and describes the results that are obtained in response to research hypothesis 4 (and its implicit questions) (RH2).

4. **Chapter 5** is drawn from the intent to meet objective 3 (O.3) and describes the research guided by research hypothesis 3 (and its implicit questions) (RH3).
Chapter 2

Literature Review

This chapter surveys the body of literature that supports the research described in this thesis. The literature covers both computational and psychology (developmental or cognitive) grounds and is divided into three parts:

Section 2.1 introduces the notions of cognitive and developmental robotics and a revolutionary perspective of cognitive robotics that implicates a strong component of human language.

Section 2.2 discusses theoretical views in which language influences human development and cognition and noted computational models of low- and high-level cognitive modelling in robotics.

Section 2.3 describes the emergence of conceptual development in humans, the notion of the Working Memory and its potential relation to cognition and draws on a key cognitive model that complies with the principles of the working memory as a pertinent skeleton to support the research conducted in this dissertation.

2.1 Robotic Paradigms

This section highlights an eminent body of literature on robotics research, narrowing the ample field of robotics in the two paradigms that fit the ambit of this thesis: Cognitive Robotics and Developmental Robotics.

2.1.1 Cognitive Robotics

The precise definition of cognitive robotics is a highly conflicted topic in literature. Many define cognitive robotics as the scientific-driven attempt to study cognition by reproducing aspects of current theoretical knowledge on human behaviour to physical artefacts (robots) (Mirolli & Parisi, 2011). Most lines of present-day research drift into this connotation. A controversial and restrictive prospect reads into using robots to study “truly cognitive” phenomena over simple embodiment for motor interactions in/with the environment (Clark & Grush, 1999, p. 12). However, this “true cognition” that involves “off-line reasoning and vicarious environmental exploration” is little researched (Mirolli & Parisi, 2011).

What is described in this section complies with the less restrictive interpretation of cognitive robotics. In describing the terminology, cognitive skills thereafter refer to all aspects of cognition that facilitate or lead to accomplishing task-oriented goals adaptively and usually in close liaison with the environment (Mirolli & Parisi, 2011). By having a physical robot enacting in the tangible world, one can study low-level cognition first and the emergence of more complex behaviours as a continuum. This low-to-high level continuum adheres to the Piagetian genetic epistemology approach (Piaget, 1968). The genesis of the theoretical basis on cognitive robotics models is affiliated with behaviour-based robotics (Brooks, 1991) and the birth of neural networks (Parisi et al., 1990). The notions around cognitive robotics were solidified in the “Artificial Life Route to
Artificial Intelligence” (Steels & Brooks, 1994), which defined true cognitive processes as “embodied”, “situated” and, to a degree, “distributed” in the environment. According to this definition, the somatic properties of the organism (i.e., embodiment), the real-world constraints (i.e., situatedness) and the interoperability with other organisms beyond the self-representation in the head (i.e., distribution) influence and are part of cognition. The concept of a physically embodied intelligent agent is encountered also in the work of Stein (Stein 1997, 471). What follows offers a more comprehensive description of the terminology, patterned after the work of Nolfi (Nolfi, 2021).

**Embodiment:** indicates entities with a physical body. The body of a robot is crafted from parts of various materials that are placed in a specific morphology. The robot’s behaviour is determined to a considerable extent by the features of its body. A definition of quantitative nature is offered by Nolfi “Embodiment indicates the extension to which the robot’s body is adapted to the other constituting elements: the brain of the robot, the environment, and the task that the robot should accomplish.” (Nolfi, 2021, 41).

**Situatedness:** a binary criterion that distinguishes systems, which are situated in and interact with an external environment, from systems that do not engage with an external environment. It can be used to illustrate to what extent a system placed in the environment can transfigure its perceptual and/or physical environment through its actions. The higher the embodiment and the situatedness of the robot, the more it will operate effectively.

**Distribution:** refers to the belief that entities (robots) do not act in isolation in their environment and that the cognitive processes do not emerge only inside the robot’s brain. They are also heavily influenced by the robot's surroundings, which may include humans, physical artefacts, and other agents.

With time, the term cognitive robotics has stretched across ample disciplines beyond cognitive sciences and robotics, such as psychology, biology, philosophy, cybernetics, given the definition of Kawamura and Browne as “the intention to model human-level intelligence in perception, motor control and high-level cognition” (Kawamura & Browne 2009, 1). Another definition assumes the notions of biologically inspired methods and autonomous life as focal attributes of a robot with motor and social cognition: “the use of bio-inspired methods for the design of sensorimotor, cognitive, and social capabilities in autonomous robots” (Seel, 2012, 613). These interpretations on cognitive phenomena and cognitive robotics seem to converge well with influential findings in theoretical psychology (e.g., Gibson, 1979; Bickhard, 1980; Norman, 1980) and cognitive science (e.g., Barsalou, 1999; Borghi et al., 2004; Hutchins, 1995). However, understanding the link between low-level competence of sensing and action with the high-level human-specific functions such as abstract reasoning, logic and creative thinking is an ultimate goal both in psychology and the cognitive robotics field. This has kindled the proposals to integrate reasoning with perception and action (De Giacomo et al. 1999), and several works that study the knowledge representation and reasoning skills of autonomous robots in only partially complete worlds (Levesque & Reiter, 1998; Levesque & Lakemeyer 2008). Interestingly, much of recent research has incorporated a revolutionary perspective to achieving complex cognitive functions in robotics: the Vygotskyan theory.

### 2.1.1.1 Vygotskyan Cognitive Robotics

Vygotsky (Vygotsky, 1962, 1978) strongly emphasises the role of language to radically transform basic cognition into spiralling level of cognitive competence that is naturally a human attribute.
Language is not simply a verbalisation of real-world phenomena. Instead, it has major implications on attention, memory, categorisation, spontaneous behaviour, and the partitioning of complex tasks in simpler steps (Vygotsky, 1962, 1978). Linguistic utterances provide cues for events that cannot be directly observed, e.g., counterfactual thinking and mediate most of the high-level human cognitive processes. The Vygotskyan perspective has spiked much interest in several disciplines, such as philosophy (Carruthers, 2002; Clark, 1998, 2006; Dennett, 1993), psychology (Gentner, 2003; Spelke, 2003; Tomasello, 2003) and proposals from the robotics field (Zlatev, 2001; Clowes & Morse, 2005; Lindblom & Ziemke, 2003; Mirolli & Parisi, 2011). They consent that language is a fundamental cognitive tool in the human developmental stages and as such could have a similar influential impact on the cognitive competence of robots.

There is a significant body of empirical and computational literature on the imprint of human language in numerous cognitive phenomena such as learning, categorisation, abstraction, and memory. These are thoroughly discussed in section 2.2.2.1. The research described in this thesis pursues the trail of ideas and methods that make use of or comply with the Vygotskyan theory of viewing language as a cognitive tool to achieve the conceptual development of artificial models and robots. With conceptual development it is meant the development of psychological representations of knowledge (Pexman, 2019).

2.1.2 Developmental Robotics

While the goal of cognitive robotics is the cognitive capabilities of robots, developmental robotics focuses on the process of forming such capabilities. Developmental robotics, otherwise known as epigenetic or ontogenetic robotics, is a scientific discipline that seeks to validate human and animal development theories via experimentation, which symbiotically feeds back novel hypotheses. A comprehensive definition of the term has been formulated and further elaborated by Cangelosi & Schlesinger (2015) as “the transdisciplinary approach of designing behavioural and cognitive skills in robots inspired by the developmental principles and mechanisms heeded in the natural child cognition”. It targets the emergence of typical skills possessed by human infants, in a lifelong and incremental manner, modelling learning as a cumulative process of progressing complexity rather than an end goal.

Developmental robotics takes its vision from schooling “baby” robots through the developmental stages of human infants, to regulate the interaction between body, brain, and environment towards achieving somatic and mental skills of ever-increasing complexity. The abecedarian constraint to acquiring such skills is the situation of the physical body of infants (humans and robots). Social learning and intrinsic motivation are also fundamental for developing cognitive functions like imitation, turn-taking, “private” or inner speech (Vygotsky, 1962, 1978). The learning results in each developmental step support the assumptions and hypotheses for the next learning stage, in nonlinear stage-like learning (Lee et al. 2007). Similar to human development, this learning is (Baldassarre & Mirolli, 2013):

1. *on-line*, i.e., it occurs in interaction with the environment and other organisms in it,
2. *cross-modal*, i.e., involves distinct cognitive domains, such as perceptual-motor and executive functions, memory, language, complex attention, and social cognition,
3. *open-ended*, i.e., it continues and evolves during the organism’s life.

In line with the above principles, the research described in this thesis follows a developmental approach, in that it is an on-line, cross-modal, and open-ended learning, taking direct inspiration from the close-to interaction of human infants with their caregivers. The coherent and
chronological method(s) described in chapters 3-5 place a great emphasis on the robot’s continuous learning of novel behaviours in/from the tangible environment (on-line, cross-modal, open-ended), majorly using human language as the natural communication interface. Unlike current advances that are unable to grow past their trained state, a major contribution of the research in this thesis is that the offered methodology to model high-level cognition enables the continuous emergence of cognitive behaviours (of developmental complexity) when the robot is “out of the factory”.

2.2 Language and Development

While advocating that the physical embodiment is instrumental for the development of manlike intelligence in robots, a paramount question emerges: what is the role of language in cognitive development? Is language merely a translation medium of our mental ideas into an interpretable communicative form or part of the cognitive thinking process? (Clark, 2006).

A prevalent perspective embraces the influence of language on thought but only considers language as a trigger to activate complex internal states or representations of the mind, which comprise the true cognitive competencies (Fodor, 1975, 1987). By learning to map linguistic encodings to the inner code representations (in the mind), the captured content can be utilised and transferred between individuals using that language code. Language is thus regarded as an innate ability or a “language of thought” (Fodor, 1975). Others underline that language has a fundamental cognitive role and is not simply a translation medium “from/to” real-world representations (Clark, 2006; Carruthers, 2002; Hermer-Vazquez, 1999; Frankish, 2007). Instead, it complements basic forms of neural processing, mapping meaning into context and coordinating context-dependent thought and reasoning. This fuels the connectionist models of memory, storage and processing (e.g., O’Brien & Opie, 2002; Elman, 1995, 2004), which stress the role of language in context representation, for instance in how novel and similar stored information items interact with one another or that information retrieval is overly context sensitive.

The research described here attends to the thought that language is not simply a public translator of our inner mental states, but a focal part of our cognition.

2.2.1 Language in Cognitive Development

What we know about the cognitive progress from infants to adults majorly stems from the legacy of Piaget on cognitive development (Piaget & Cook, 1952). The Piagetian theory neglects language eminently when describing four major stages of development: sensorimotor (newborn up to 18-24 months), in which infants discover their surroundings through senses and develop fine motor skills; pre-operational (ages 2 to 7), during which symbolic thinking emerges and the ability to represent concrete phenomena into words; concrete operational (ages 7 to 11), in which children demonstrate inductive logic and reasoning and can recognise a difference in their thinking with respect to others; formal operational (age greater than 11), characterised by highly abstract thinking and reasoning hypothetically. This theory is majorly contradicted by Vygotsky, which foresees no stages of development but only dependence on social context.

Most works that have developed in response to Piaget’s theory either in favour (e.g., Fischer, 1980) or opposition of (e.g., Spelke, 1991) seem to sustain the thought that cognition and language are independent of one another. Differently, Nelson (1998) advocates that language rests on prior or innate pre-linguistic, conceptual and social processes and as such is not separable from cognitive development; in turn, cognitive development is not separable from language. For instance, it has
been demonstrated that the human ability to sort items can emerge in the absence of language, but this does not seem to always match the sorting done in response to language (Malt et al., 1999). In the initial stages of development, infants attempt to make sense of their world by forming representations of objects, events, and relations, to which they associate linguistic categories to express thought. These mental representations of real-world referents can be thought of as universals in language or language-independent (Clark, 1983). During the first year of life, children learn to encode their experiences in words and form a concrete item-based vocabulary of object-word mappings; however, over the course of their development (many years), they become exposed to and begin to self-construct more abstract high-complexity linguistic constructions, which lead to non-trivial reasoning, and thus, challenging the hypothesis of whether language is innate or not (Tomasello, 2000). This conundrum becomes more evident when comparing the conceptual domain of children that speak different languages. While the primate stages of language development seem to not differ much when and once languages are learned, the way ideas are evoked and how experiences are described varies from one language to the other. Each language uses words in its way to draw attention to some aspects and ignore others, or to pick out relevant categories. That means, the more the linguistic complexity increases with time, the more the conceptual representations will diverge between the different tongues. Older infants will map different linguistic constructions to the same conceptual domain, thus how they talk about each domain will majorly reflect the language they speak or think in (Ghent, 1961).

There are two opposing perspectives to whether language plays a role in our knowledge representation or simply labels this knowledge to communicate it. When the symbolic capacity to map words to concepts emerges in the pre-operational stage of Piaget, does that mean linguistic labels map onto concepts that are already formed independently (word-as-mapping view)? This perspective, also known as the “cognitive priority” hypothesis (Bowerman, 2000) paints a rather insignificant role of language in the acquisition of knowledge. However, it brings forwards at least two limitations. One limitation is that it cannot explain the context-rich meaning of words, e.g., to how many concepts does the verb TAKE map onto in the following “take a picture”, “take a breath”, “take the bus”, …, which differ in meaning, yet share a similarity in the various senses of “take” (Lupyan, 2019). A second limitation becomes more evident in the cross-linguistic labelling of concepts/patterns. To express that something is hard to understand, in English, you would say “That is all Greek to me” (that roots to Latin), whereas in Greek one would say “I see it all Chinese”. Though the meaning might be similar, it would seem that the concept of something unintelligible and gibberish for English-speaking people relates to Greek symbols/alphabet, while to Greeks this is mapped onto the concept of Chinese structures. The alternative to this perspective is the one that regards word-as-cues, by which language is afforded a cardinal role in creating and augmenting the type of knowledge acquired through the motor and perceptual experience (e.g., Elman, 2004, 2009; Lupyan, 2016). In this view, words cue to meaning and aid the development of our conceptual repertoire, placing language into equal share with our direct perceptual and motor experiences in the process of cognitive development.

It is important to highlight that the research described in this thesis focuses not on the stages of language development. Rather, this research heeds the salience of language in the development and amelioration of human cognition, by taking a words-as-cue perspective and, as such, it suggests that language can be an effective tool to support or enhance cognition in artificial artefacts (e.g., robots).
2.2.2 Computational and Robotic Language Models

The cardinal aim of this thesis is to model aspects of high-level cognition (section 2.2.2.1) for cognitive robotics, relying on the belief that human language is a pivotal cognitive tool to achieve this target. Its motivation stems especially from the fact that the current body of literature (section 2.2.2.2) despite offering significant advances is limited to low-order cognitive phenomena, which can, however, be used as a cornerstone to building on more complex cognitive modelling. As language is a keyword of this thesis, it is worth surveying the current efforts to model its acquisition and comprehension (sections 2.2.2.3, 2.2.3), to draw on a method that exploits language in a cognitively plausible manner to model conceptual behaviour.

2.2.2.1 Models of high-level cognition

This section reviews seminal literature on computational and robotic language models, which have attempted to or have addressed high-level cognitive phenomena, majorly focusing on learning, categorisation, abstraction, and voluntary control.

- **Learning**
  
  It is argued that language has a profound impact on learning, in that it facilitates the learning and formation of categories: (a) by directing our attention onto the meaningful stimuli from the environment that are relevant for categorisation; (b) by granting hints to categorise those stimuli (often only a language attribute), which helps develop the discriminative capacities required for categorisation. Assorted computational models have studied how linguistic labelling can speed up category learning, e.g., of visually presented entities (Schyns, 1991) and improve their internal mental representation (Lupyan, 2005). Steels & Belpaeme (2005) have demonstrated how the conceptual system of a collection of autonomous agents adapts itself while acquiring a self-organised linguistic repertoire of perceptually grounded categories, to optimise their cooperative communication. Further work on artificial life organisms that use language for the adaptive categorisation of their experience has considered the influence of social learning (Cangelosi et al., 2000, Cangelosi & Harnad, 2000). Their findings suggest that learning through social cues first requires an autogenous capability to categorise some entities and next, an appropriate exposure to linguistic stimuli from other organisms in the environment to categorise novel experiences. This approach has been shown to speed up learning and adaptation, while minimising the cost of error, compared to other methods.

- **Categorisation**
  
  Language can enhance categorisation after categories have been formed or learned. A noteworthy neural network model that has attempted to address this cognitive competence can be retrieved from the work of Mirolli & Parisi, (2005), (2006). The authors modelled the interaction between linguistic labels with sensorimotor activation in a modular neural network that simulated the “child’s brain”. The model was trained in stages similar to child development to label perceived categories (name objects) and to discern perceived words (produce an appropriate sensorimotor response). From the neural activation patterns, categories were determined as the activation set elicited by sensory input, which generated the same motor response. For optimum categorisation, objects belonging to the same category should fall as close as possible to the activation pattern of that category and as distant as possible from other categories. The model demonstrated that the internal representations of pre-linguistically learned objects would change when the model learned to map those objects with their respective linguistic labels. By introducing linguistic inputs, the objects would fall more closely to one activation pattern and more distant to other activation patterns, thus enhancing their categorisation. These findings suggested that organisms with
language display superior behaviour, making the model the first computational model which explicitly substantiates the Vygotskyan postulate.

- **Abstraction and Voluntary Control**

  When categorising different stimuli for which the expected response is similar onto the same category, their dissimilarities should be critically dismissed, i.e., *abstracting from their differences*. To categorise similar stimuli which elicit different expected responses into separate categories, their similarities should be neglected, i.e., *abstracting from their similarities*. This kind of abstraction cannot exist without context. Language encourages abstraction by abolishing the irrelevant characteristics of an entity for a certain context and preserving only the essential characteristics that fit pragmatically in that context.

  A computational model that draws on the Vygotskyan idea of inner speech to empower cognitive functionalities such as categorisation and abstraction is introduced by (Granato et al., 2020). The model is tasked with solving the Wisconsin Card Sorting Test (WCST), in which a deck of 64 cards must be sorted based on 4 pre-set target cards. The deck involves a combination of geometrical shapes (stars, triangles, circles, crosses) appearing in varying colours (red, green, blue, yellow) and numbers (one, two, three, four). Each target card has a unique set of attributes that are not shared with the other three, e.g., card 1: four blue circles; card 2: three yellow crosses; card 3: two green stars; card 4: one red triangle. The sorting criterion is unknown (shape-based, colour-based, or number-based). The model attempts to sort the card by selecting an arbitrary criterion. The experimenter gives language feedback (“correct” or “not correct”) after each sorting attempt, which prompts the model to adapt to the inferred rule given the verbal feedback. The model was able to reproduce human collected data on the WCST task, verifying many empirical findings, which suggest that the categorisation process can generate more abstract patterns recursively (Mirolli & Parisi, 2011). For instance, initially sorting the cards based on colour requires abstracting from their other attributes such as number or shape. When inner-speech feedback is introduced, the attention may be drawn away from colour and drawn into another attribute, say shape, therefore generating abstract concepts progressively like “colour”, “shape”, “number”.

  On the other hand, this study seems to validate how the internalisation of linguistic stimulation (inner speech) shapes the problem-solving decisions of the model to adapt its behaviour favourably when relevant stimuli are introduced. i.e., voluntary control. Voluntary control is a consequence of the cogent skill of abstraction and in the discussed model it arises from the linguistic stimulation feedback (“correct” or “not correct”), which triggers the model to modify its actions. The research is in line with empirical studies that have addressed how talking to oneself fundamentally impacts goal-directed behaviour (Boysen et al., 1996).

### 2.2.2.2 Robots that use language

The benefit of language (overt form) and inner speech (covert form) to empower cognition requires a sufficing symbolic capacity to associate internal mental representations of concrete phenomena with linguistic labels (Mirolli & Parisi, 2011) (Granato et al., 2020). Theoretical and empirical evidence from developmental and cognitive psychology, cognitive linguistics and neuroscience claim that embodiment and mastery of language grounding are required for cognition (Wilson, 2002; Pecher & Zwaan, 2005). Noted methodologies of computational modelling of language grounding in developmental robotics (Cangelosi, 2010) and integrative models of motor and cognitive capabilities (Steels, 2003; Feldman & Narayan, 2004; Perlovsky, 2009) have positioned themselves in favour of validating those findings via experimental deployment in artificial agents. This thesis also adheres to the embodied approach of integrating language with
Chapter 2: Literature Review

The literature below surveys models for learning language meanings, which is of particular interest for the proposed learning methodology described in Chapter 4 of this thesis.

Grounding language to perception: Mapping perceived objects from images or real workspaces with their linguistic labels is an area of extensive research (Glenberg & Kaschak, 2002; Borghi et al., 2011). Deep learning approaches using deep convolutional neural networks have achieved ground-breaking results on recognition tasks, like labelling an object class in an image (Krizhevsky et al., 2012) or dense recognition with natural language description of all or salient objects within an image (Johnson et al., 2016). These methods have been enacted in robots to train the skill of object retrieval from human instructions using simulated images (Hatori et al., 2018), followed by attempts to study how well can (a fixed vocabulary of) vision-language mappings work in a real-world robotic setup, using the iCub robot (Štepánová et al., 2018). Language has been exhaustively used in robotics research for object referencing, of their features (colour, shape, number) or in-detail description of the object type, category, and other high-level attributes, to improve robot learning (see Tellex et al. 2020 for a review).

Grounding language to motor representation: Linking sensorimotor representations and natural language conform to the notion of embodied cognition has been extensively examined in the literature. Numerous works tackle the problem of language acquisition and grounding for use in engineering contexts. In probabilistic models, symbols are derived from raw data, including speech (Iwahashi, 2008), video and motion to learn visually grounded vocabulary (Roy & Pentland, 2002) and the meaning of manipulation verbs (Roy, 2003). The instruction following using real robots for object manipulation in response to human language has been well modelled as a reinforcement learning (RL) problem (Levine et al., 2016; Gu et al., 2017). On the pitfall, RL techniques are designed around highly specific robot functionalities and do not offer a robust solutions to general problem-solving. The learning process is poorly understood, and, in most cases, the underlying mechanisms of language acquisition are neglected. To boot, seminal contributions on integrating language with the physical body are modelled as ad-hoc solutions and at-random exploration of language for highly constricted tasks (Morse & Cangelosi, 2017; Demiris & Khadhouri, 2006). Pivotal results arrive as well from parametric Bayesian modelling in a comprehensive multi-modal framework that jointly addresses word learning, knowledge acquisition and decision-making, but not without the cost of a pre-defined number of classes, which impedes open-ended learning Miyazawa et al. (2019).

A prevalent artifice to self-organise language-meaning representations in robotic architectures is the use of supervised learning methods, amid which some have taken significant steps towards humanlike intelligence demonstrated in learning experiments with real robots (Ogata et al., 2007; Yamada et al., 2016; Tani, 2016). For instance, Yamada et al., (2016) devised a complex representation learning approach to language-motor mappings and the grounding of logic words ("not", "and", "or"), a compelling area underexplored in most conventional works (Yamada et al., 2017). However, these proposals mainly implicate low-level motor skills and in a way that neglects perspectives from cognitive linguistics and psychology. By contrast, this has inspired the efforts of several authors that model usage-based language acquisition and production, unidirectionally (Sugita & Tani, 2005; Hinaut & Wermter, 2014; Hinaut & Twiefel, 2019; Moulin-Frier et al, 2017) or bidirectionally (Heinrich et al., 2020), i.e., learn motor meaning from language and emerge language skills from motor exploration.

Language-driven learning in the wild: A key aspect that has been generally absent in the body of literature surveyed so far is the ability to learn incrementally and offline, to improve the robots’ knowledge about the world. Large corpora of training data, pre-defined vocabularies and
behaviours or ad-hoc learning policies cannot anticipate real-world endeavours and/or require augmented computational effort. Authors have attempted to mitigate this gap using, for instance, dialogue-based methods, in which a human guides the robot to execute new instructions it has not heard before, by explaining it in simple terms that are broken down into step-by-step descriptions (Cantrell et al., 2011). The merit of learning from natural language descriptions has also been validated in the generation and enhancement of the semantic representation model of the environment (Walter et al., 2013). Moreover, it has been demonstrated that a robot can begin with limited knowledge of the meaning of words and gradually develop language-visual mappings to parse commands about previously unseen instances through cognitively plausible modelling of grammar learning (Alomari et al., 2017).

A richer array of works beyond those surveyed in this section can be retrieved from Tellex et al. (2020). Nonetheless, much of the body of research, be that highly language-driven or not, has been highly focused on elementary lower-order cognitive phenomena (perception, manipulation, motor coordination, navigation) and there is no immediate indication of how this might scale up to explain complex human cognition. Moreover, many of these advances rely on deep learning. It is fair to say that deep learning methods have propelled a broad range of domains, with a great upswing in artificial neural networks and connectionist architectures being used to address the hindrances of embodied language processing. However, as a final note it is worth listing some typical limitations, especially when extending them to robotics (Pierson & Gashler, 2017):

i. Deep-learning methods are data-hungry, needing huge learning data to achieve satisfactory performance. The generalisation of the test data often does not represent the generalisation under real-world constraints. Moreover, learning from large corpora is not the intuitive way humans learn.

ii. They are habitually designed and optimised around a specific ad-hoc problem, for which goals and rewards might change only slightly. They are unable to generalise across situated contexts and tasks, significantly limiting their ability to adapt in the wild. The environment and task are often too constrained and while this allows learning more efficiently from less data, it drastically inhibits the generalisation competencies of these models.

iii. Despite their access to greater amounts of learning data, this background is much restricted compared to the distribution of data encountered during a lifetime. Most deep-learning methods fail to acquire new data and apply them directly in new constraints, as they focus on input-output approximations, instead of modelling action-inference associations.

### 2.2.2.3 Other language processing domains

While this thesis conforms to an embodied approach and heeds human language as a tool to model cognition rather than language modelling per se, it is worth reviewing major advances of neural network language models in other domains and their adequacy with the perspective of this thesis. This section surveys pre-trained models and cognitive architectures intended for language processing and development and emphasise key differences with the research proposed here.

- **Pre-trained models**

The effort to produce and comprehend human language has led to prodigious research on Natural Language Processing (NLP). These techniques train disembodied neural networks to develop broad language functionalities, such as text translation, summarisation, paraphrasing, question answering or describing the content of images and videos using phrases of human language. The field has advanced from hand-crafted and statistical methods to deep neural networks that self-
learn task-specific features from raw data. This comes at the cost of huge data requirements to avoid overfitting and improve generalisation (Belkin et al., 2019), which has led to the creation of vast corpora and has been supported by the substantial boost of computing resources. We can see this reflected in the commercial deployment of human language technology such as conversational personal assistants (e.g., Siri, Google Assistant, IBM Watson) or standalone devices (e.g., Amazon Echo, Google Home/Mini).

The current state-of-the-art in NLP is closely affiliated with large neural language models such as BERT (Devlin et al., 2019) and GTP-2 (Radford et al., 2019). These models have demonstrated that they can learn information about linguistic phenomena like subject-verb agreement (Talmor et al., 2020; Jawahar et al., 2019), and semantic role types from context (Tenney et al., 2019), all trained in English. Much of NLP research claim great successes on meaning-sensitive tasks (Habernal et al., 2018, Williams et al., 2018). However, careful studies have found that these models were simply more dextrous with data artifacts but far from reasoning on the task, performing much below chance when the datasets they are trained on are frustrated in some way (McCoy et al., 2019). There is still a long way to understanding in-depth what aspects of human language these large models represent (Bender & Koller, 2020). While they are originally inspired by the complex hierarchical organisation of the cerebral cortex in the brain, they have gradually evolved into empirical engineering approaches to solve very specific problems. Therefore, classical NLP techniques of language development fail to cater for an empirically meaningful computational model of cognition that can provide intelligible insights into how the human brain acquires, comprehends, and produces language (Dupoux, 2018; Kelly & Reitter, 2018).

Furthermore, even with the upsurge of large-scale pre-trained models (PTM) (Han et al., 2021) that can capture knowledge from huge corpora of both labelled and unlabelled data, their linguistic competence is constrained due to the limited interaction modalities (Nolfi, 2021). The ability to describe observed objects is insufficient to grasp the rich meaning that arises from the tactile interaction with the objects. For instance, recognising a chair by its canonical features and simply knowing that you can sit on a chair (observable) cannot pragmatically categorise other sitting objects as chairs if lacking the experience of sitting on a chair (tangible). Consequently, the type of language modelling offered by these systems cannot be effectively used to support cognition.

### Cognitive architectures

The surveyed works are narrowed down to neural systems that model brain processes and cognitive architectures that avail from the principles of the Working Memory (WM) (Baddeley & Hitch, 1974; Cowan, 1995). The term cognitive architectures here refers to relatively large computational models composed of heterogeneous parts and subcomponents, intended to simulate aspects of human cognition and behaviour (Ye et al., 2018). Studies suggest that the emergence of language and its processing competencies can be explained in large part in terms of working memory operations (Baddeley, 2010). Implicitly, there is a level of language processing that is independent of language and involves the flow of information between working memory buffers (Verhagen & Leseman, 2016). Noted architectures that attend to this principle include ACT-R for intelligent behaviour (Lovett et al., 1999), Soar (Laird et al., 1987) used to model perception and action, the EPAM general architecture extensively used in experimental mental paradigms (Feigenbaum & Simon, 1984), the combined framework of symbolic memory and emergent attention for human cognition (COGNET/iGEN) (Zachary et al., 2000) and the ANNABELL large-scale brain-inspired cognitive architecture used here (Golosio et al., 2015). A more encyclopaedic review of cognitive architecture research in broader domains can be retrieved from (Ye et al., 2018). The aim is not to survey these models in-depth, but to argue that modelling linguistic competencies (i.e., knowledge
about language) in cognitive architectures may allow the emergence of linguistic performance (i.e., its application in concrete endeavours) (Rij et al., 2010) and successful modelling of human cognition for robust problem-solving in robotics (Kurup & Lebiere, 2012). They allow their generated predictions to be probed as attested by empirical data and findings. While realistically there is a possibility that cognitively informed models may apply different or even inconsistent assumptions with the linguistic theories, given their limitations compared to the human brain or other simplifications, they are a promising tool to understanding the issues that underpin language domains and test linguistic competencies empirically. They stem from well-established theories of human cognition and allow to study linguistic phenomena under variable constraints. This upholds why the research described in this thesis exploits a brain-inspired cognitive architecture to validate its hypotheses.

2.2.3 Multilingual Computational Models

This section reviews the literature on multilingual models. Multilingual computational models, in particular those that simulate aspects of the human brain (e.g., cognitive architectures), can be used to study focal phenomena of bi- and multilingualism. Suitable cognitive modelling of poly languages can be beneficial to debrief the empirical findings in the simultaneous acquisition of languages and the “tuning” observed in the baby brain to reflect the languages it encounters from birth (Crinion et al., 2006). From a non-cognitive angle, the increased commercial interest in multilingual services and households has resulted from a growing number of polyglot speakers outnumbering their monolingual counterparts. This thesis offers not a computational linguistic modelling as its prime goal but models poly language acquisition in a way that is cognitively plausible with an eye toward contextualisation. Contextualisation can lead to the understanding of context and cultures by grasping the language’s intricacies specific to that context/culture (Hassan, 2014).

Given the current eminence of computational psycholinguistic models of monolingual acquisition, the near-absence of computational bi-/multilingual modelling is striking. The desiderata for such models that could explain empirical findings have been defined as follows (Frank, 2021):

- Simulate properties of language processing in at least two languages, factoring their disparities in proficiency and exposure.
- Cater for relevant empirical data of bi-/multilingual processing phenomena, such as code-mixing/switching, i.e., alternating between languages in a conversation, crosslinguistic transfer and priming, as observed in humans.
- Be cognitively and/or neurobiologically plausible without unreasonably augmenting the complexity, for example, the capacity of the long-term memory, with respect to a monolingual model.

Bilingual models: Recurring practices to construct bilingual models are by training “parallel corpora” of translation equivalent sentence pairs from both languages (Burkett & Klein, 2008; Saers et al., 2012) or bilingual corpora that do not share the same content among the languages (Cohen et al., 2011; Iwata et al., 2010). These models offer some useful insights on an adult bilingual’s language knowledge, but they do not form a promising cognitive model that could describe the natural way humans learn multiple languages (the former being less cognitive than the latter) (Frank, 2021). As such, they can be a technological solution vaguely supporting any empirical findings. Their utmost drawback is the augmented complexity and structural changes that are required to elaborate more than one language. These models do not assume a monolingual system that learns a new language but necessitate the learning of explicit links between grammar...
constructions that are translation equivalents between the languages (e.g., Burkett & Klein, 2008). The use of Bilingual Recurrent Neural Networks (RNNs) has mitigated the stipulation of architectural changes. Such models can self-learn the languages and their identity using word co-occurrence, which is more prominent in one language than across languages, to carry out tasks like next-word prediction (e.g., French, 1998; Frank et al., 2016). Nevertheless, these models are habitually designed for certain technological applications such as automatic translation over cognitive modelling. Bilingual RNNs have shown promising potential for bilingual sentence comprehension. Trained on a small-scale bilingual corpus in English and French, the model of Hinaut et al., (2015) could learn the languages together with similar performance compared to learning them one at a time, without explicitly defining their identity at the input. When applied to robotics, the model could successfully elaborate 15 languages, using a miniature human-robot interaction corpus (Hinaut & Twiefel, 2019). This model raises important assumptions, which suggest that RRN could in principle be beneficial for the cognitive modelling of bilingual comprehension. However, extensive research and testing on human data is indispensable to understanding what about these models is relevant to bilingualism research.

**Multilingual models:** Current literature has led to the epiphany that the transition from bilingual to multilingual models is burdenless and straightforward; however, these insights have emerged from the computational linguistics domain and are little or not tackled in cognitive models, which can be closer to the empirical frameworks of human multilingualism (Frank, 2021). While there are notable multilingual models of grammar induction, which demonstrate the successful learning of four (Cohen et al., 2011) or more languages - ten (Iwata et al., 2010), without supervision or parallel corpora, their reliance on probabilistic grammar modelling fastens these solutions to specific architectures to handle multiple languages at once, posing, in addition, a significant uncertainty on their psycholinguistic relevance (Frank, 2021). In contrast, vast multilingual models established on neural networks require little to no architectural adaptation relative to the number of the learned languages, demonstrating significant potential in translation, including focal aspects of code-switching (e.g., Johnson et al., 2017). Though performance might degrade when the corpora are augmented with more languages, multilingual models have often demonstrated a significant boost in the learning performance of those languages for which the training data is limited. Large neural network language models (LMs), such as multilingual BERT (Devlin et al., 2019), can learn features of linguistic formal structure, but this ability is extensively built on leveraging artifacts of the learning data and represents little analogy with human language acquisition (Bender & Koller, 2020).

It is in the belief of this thesis that cognitive architectures, which adhere to the principles of the working memory (WM), can capture important aspects of language development and may drive intelligible conclusions about the human brain mechanisms involved in the acquisition, comprehension, and production of multiple languages. This motivates why the research described here first draws on a WM-compliant framework for the suitable cognitive modelling of multiple human languages agreeing with the three desiderata for bi/multilingual models (Chapter 3).

### 2.3 What is Conceptual Development

#### 2.3.1 Concepts and their Understanding

*What is a concept?*

In the simplest of definitions, concepts are notions we use to represent categories of either material or abstract existence (Murphy & Medin, 1985). They consist of a range of attributes, i.e.,
characteristic qualities: some are critical to the definition of a concept, while non-critical attributes can be variable, yet do not define the concept. Whether attributes are critical or not varies from instance to instance: colour may be a typical but non-defining attribute of the concept of “cat”, but it helps separate “water” from “milk”. Certain human intellectual processes, such as planning, thinking, reasoning, problem-solving, and decision making, rely heavily on concepts. The way humans acquire concepts and employ them in their thinking in/ about the real world is central to their learning and cognitive development (Sloutsky and Deng, 2019).

According to Bruner & Austin, (1986), concept learning typically involves two major tasks. Concept formation refers to the attempt of sorting an instance into a meaningful class. Concept attainment is the process of assessing the significant attributes that distinguish exemplars from non-exemplars of the class. It entails the governing processes involved with categorisation. Grouping concepts pragmatically into coherent categories allows us to draw non-trivial inferences on situations in which we lack direct experience. Categorisation as a cognitive ability leads to decision-making, i.e., doing something with a befitting kind of thing.

Traditionally, cognitive linguistics suggested that a concept has a definitional structure of necessary and jointly sufficient conditions that classifies that concept unequivocally under a specific category (classical theory) (Taylor, 2003). New complex concepts can then be formed by combining simpler definitions, and candidate concepts are categorised by verifying that all and every attribute fits the constitutional definition. However, concepts are generally not well-defined. Most concepts do not have precise definitions and can be convoluted or atypical, which has led to controversial findings assuming that concepts are organised around a typical example of the concept, either a prototype (prototype theory) (Rosch & Mervis, 1975) or common category representatives (exemplar theory) (Medin & Schaffer, 1978). These suggest that concepts have a prototype structure (probabilistic) and that category members are assessed given their degree of similitude with the prototype or the exemplars of the category. Concepts of the same category show some family resemblance (Rosch & Mervis, 1975; Wittgenstein, 1953), of typical entities that possess attributes that are most common in that category and less frequently found in other categories. A candidate concept must possess enough significant attributes to be accepted as a category member.

What does it mean to understand a concept?

Being able to formulate an explicit definition of a concept is not required for (Bruner & Austin, 1986) and does not necessarily indicate (Skinner, 1957) understanding. The understanding of the concept is guided by its critical (defining) attributes over its non-critical (variable) attributes (Danahoe & Palmer, 2009). Classifying an instance pragmatically in a group is part of problem-solving. Ergo, the first step to understanding is categorisation. Categorisation helps reduce the complexity of the environment and the necessity of constant learning. If an entity belongs to a certain category, it can be used in close contexts as the other members and produce similar interactions with further categories, needing not to learn each member de novo. Significant research has demonstrated concept/category learning at a human level, even on creative generalisation (Lake et al., 2015; Lazaro-Gredilla, 2019). Yet, these advances have majorly considered perceptual categories (vision/action-based) without accounting for natural language that grants progressive levels of abstraction. Language is a remarkable cognitive ability, which would enable machines to learn high-level concepts and greatly improves categorisation once categories are pragmatically learned. The conundrum faced in the current body of literature on robots that use language (see section 2.2.2.2 and Tellex et al., 2020 for an extended review) is their
ability to acquire humanlike high-level cognitive competencies beyond mere somatic skills and our ability to affirm with certainty if they are, in truth, understanding concepts evoked in language. From a psychological and philosophical angle, the learning sciences research (Sawyer, 2005) have come to consensus that understanding is demonstrated if the learner can:

C1: Identify examples of the concept that are subject to a high variation of non-defining attributes.
C2: Distinguish exemplars (an example of the concept) from close non-exemplars (something that is not an example) by assessing their significant attributes.
C3: Maintain these abilities in novel endeavours that were not presented when learning the concept.

This definition feeds the research conducted in chapter 5 and the focal hypotheses of conceptual development raised in this thesis (see 1.3).

2.3.2 The Path to Conceptual Development

Concepts might emerge either through encounters with the environment first and expressed in words later (i.e., get lexicalised) (Markman, 1989), or from language first and then anchored in the environment (Vygotsky, 1962). The discussion on the role of language in conceptual development described in this section stems majorly from the work of Sloutsky & Deng (2019). They illustrate this with the following example: infants can create a category for dogs after many encounters with dogs, but before learning the linguistic term dog. Instead, notions such as germ cannot originate directly in experience, but in language and their meaning is grounded later into some form of personal experience. Bottom-up concept learning (from experience) occurs without language, whereas top-down learning (from language) requires a fair amount of language. What we know about the human brain is that it makes use of the notion of sameness or equivalence. We tend to find patterns that allow us to treat different entities as if they were similar in some way. When we establish an equivalence (category), we can lexicalise it, and use the label to mark those entities equally. Ergo, the label becomes a concept. The process of development can progress both from category formation to its lexicalisation or from a lexical term to the formation of a category.

From Categories to Concepts: Pre-language

Things of material existence in the world can share at least one essential observable feature or an array of overlapping similarities with one another. For example, cats are identifiable by their soft fur, sharp claws, large eyes, short snout, triangular ears, flexible body, and a long slender tail. This category can be explicitly learned through observation alone and without language. The type of categorisation that relies purely upon perceptual properties is known as perceptual grouping and constitutes the simplest form of conceptual behaviour. Such categories can be extended by dint of global familiarity, with other instances that share enough salient features with category members.

Perceptual grouping can become significantly complex when mutually exclusive categories must be formed at the same time (Sloutsky & Deng, 2019). When separating cats from dogs, with which we are equally familiar (i.e., exposed to equal encounters of), the cognitive burden of the categorisation decision increases, as more categories must be processed simultaneously based on their respective observed features, which can often overlap considerably (e.g., both cats and dogs have four legs, fur and can be of the same size). However, the purpose of categorisation is not to simply group entities into equivalence classes, but to use categories in reasoning, decision-making, prediction, and inference. This competence is a further complicated form of conceptual behaviour,

---

2 C1-C3 refer to criteria 1-3 of understanding.
which can be facilitated by lexicalising categories (i.e., marking the category with a linguistic label). Associating categories with lexical terms (i.e., lexicalisation) allows gathering further knowledge about the category, which is not directly observable, for instance through social interaction (I took the dog to the vet for examination). In this way, humans learn to form convoluted category networks given the correlation that occurs among distinct categories. For instance, they may associate dogs with vets or dogs with bones, which stretches beyond simple perceptual grouping to complex conceptual networks that are likely a unique attribute of human cognition.

From Concepts to Categories: Post-language

Although language is not a necessary entry for the formation of certain perceptual categories, it can, as explained above, enhance the conceptual development of the learned categories, by adding knowledge about unobservable features of those categories. On a further note, some notions may be purely unobservable or abstract (heat, peace, germs), for which linguistic input is fundamental if not the only way to learn them. When a category is associated with a linguistic label, the label becomes part of its representation and is used to communicate this category with others or connect new information to it. Such information emerges in the form of verbal description, for which lexicalisation is critical: the statement cats are domestic animals cannot be understood by observation, if the concept domestic is not known. This categorical convolution and hierarchy formation is a product of development (Sloutsky & Deng, 2019).

Learning words for categories contribute to the formation of the type of knowledge known as semantic memory, which is focal to our capability to use concepts in high-level cognitive phenomena, such as planning, explanation, reasoning, prediction and problem-solving. Children may learn category lexicalisations at the same time as learning the perceptual category. This occurs when the child is verbally dictated the label of a referent explicitly while showing or pointing at the referent. However, many studies suggest that a good number of words are learned from context (Goodman et al., 1998; Nagy et al., 1985), often without the explicit presence of the referent. How the meaning of words is learned from context has been an extensive area of research, birthing several proposals. While it is not central to revisit these proposals in-depth, it is worth noting that their findings suggest that many concepts emerge solely from language, as context, in particular that in which the referent is not present, involves numerous unobservable phenomena. Moreover, there is evidence that with the growth in language during child development, emerging concepts for both known and unknown categories are arranged into some semantic memory or conceptual network (Rogers & McClelland, 2004), which represents the individual’s knowledge about the world. In a nutshell, language constitutes a rich source of knowledge about concepts that complements and enhances the knowledge acquired from perceptual experience, leading to the formation of complex conceptual behaviour.

2.3.3 Embodied Conceptual Development

The previous section discoursed how concepts develop from category formation and progress to more complex conceptual networks through language. To recollect, this thesis draws on the claims of embodied intelligence. Intuitively, one may ask: what is the role of embodiment in conceptual development?

Robust research focuses on the role of sensorimotor models in the formation of human concepts. Concepts are commonly regarded as bodies of knowledge stored in our long-term memory that are used in high cognitive activities (Machery, 2009). Conceptual representation involves motor skills, perception, emotions, and language (Hargreaves & Pexman, 2014). Notably, strong embodiment
theories (Glenberg and Gallese, 2012) assume that manipulation experience is fundamental for the representation of concepts, as it is predicted that infants learn linguistic labels easier when they have action experience on it. However, while there is much consensus on the importance of motor information in conceptual development during early infancy, the predictions are less clear for older children and adults (Pexman, 2019). Research that complies with the embodied account predicts that acting in the environment facilitates development greater than observation, observed for instance in the learning of verbs (Wellsby & Pexman, 2014; Hald et al., 2015). However, embodied cognition is yet not fully able to explain how notions with poorer sensorimotor information that cannot be experienced through senses can be grounded to their meaning (Pexman, 2019). Addressing this has led to at least three proposals. The first two, grounding through perceptual metaphors (Gibbs, 2006) and grounding through emotions (Barsalou & Wiemer-Hastings, 2005; Hansen et al., 2012), are not discussed here. The third offered proposal assumes that abstract concepts are developed through language (e.g., Andrews & Vigliocco, 2010; Antonucci & Alt, 2011; Borghi et al., 2011; Clark, 1998). In particular, it has been observed that adults’ learning of new abstract notions is significantly facilitated by verbal explanations, whereas that of novel concrete notions is not (Borghi et al., 2011). Given that adults have solid language knowledge, the conclusions drawn from the study were that language acquisition influences the representation of concepts. This proposal reconciles less with the strong embodiment theory and rather leans towards the weak/weaker embodiment developmental theories: Dove (2011) – language enables the formation of new conceptual representations in infants (linguistic perceptual symbols), which are helpful for the representation of abstract concepts – and Howell et al. (2005) – child learning starts with concrete concepts directly grounded in experience and progresses to abstract concepts grounded by linguistic experience as well as by their relationship to earlier-learned concepts.

The research described in this thesis is consistent with the weakly embodied developmental theories duo. The ideas offered by these theories are particularly useful to draw the methodology in Chapter 4, by which a) the model generates self-representations for abstract words by learning linguistic perceptual symbols for concepts (Dove, 2011), b) evolving from concrete concepts with direct tangible experience to the indirectly grounded meaning of abstract concepts by representing them in relation with earlier-learned language primitives (Howell et al., 2005) (see Chapter 4).

### 2.3.4 Working Memory and Conceptual Development

Many influential works in psychology suggest that human behaviour and action control can be largely favoured by language (Vygotsky, 1962; Luria, 1959; Baddeley et al., 2001). In turn, language comprehension is fundamentally influenced by cognitive phenomena and the ability to maintain and manipulate information of verbal and visuospatial nature (Caplan & Waters, 1999; Acheson & MacDonald, 2009). These theoretical and empirical findings have prompted the omnipresent idea of working memory (WM). The working memory is assumed to be the cauldron of concept formation in cognitive psychology. In simple terms, the WM is used to describe the small amount of readily accessible information held in the mind that is used in complex cognitive tasks (Baddeley, 1992). Two major types of memory unanimously recognised by the majority of prevalent approaches on the organisation of information in the human brain are short-term memory (STM) and long-term memory (LTM). The STM can be defined as a unitary system, able to retain a small but sufficient amount of information over a brief period of time, while the LTM holds the vast amount of information acquired during one’s life. Instead, the working memory (WM) includes, as agreed by most researchers, a short-term storage and the procedural mechanisms used to collect and manipulate the information stored in it (Cowan, 2014). There is much agreement on the potential relation of the WM to concept learning (Baddeley & Hitch, 1974). Concepts are
presumably formed first in the working memory and are permanently transferred to long-term memory, where they are stored over an extended period. The working memory is amply associated with reasoning, decision-making, executive function, information elaboration, problem-solving and learning (Cowan, 2014). It is a prevalent idea in cognitive psychology, which underpins how humans learn, and could potentially influence how artificial models learn.

2.3.4.1 A Working Memory Compliant Framework

This thesis leverages an existing cognitive architecture as its framework (Golosio et al., 2015). What qualifies the architecture as a cogent skeleton to draw on research that addresses the pitched hypotheses is its consistency with the well-defined theoretical multicompartment Working Memory (WM) model (Baddeley & Hitch, 1974; Cowan, 1995). The procedural implication and analogy with the working memory drive cognitive learning in a way that is closer to the mechanisms of information elaboration and reasoning in the human brain. This promotes the capacity to learn from a limited sample of training data and generalise on ample corpora, using language to govern action-inference. As such, it accommodates the means to explore the significance of language on the humanlike development of higher-level cognitive phenomena in robots compatible with findings in psychology on human behaviour. This section briefly summarises the design and prominent aspects of the cognitive neural architecture (ANNABELL - Artificial Neural Network with Adaptive Behaviour Exploited for Language Learning).

ANNABELL is a large-scale computationally efficient artificial neural network designed to learn through a child-like developmental approach relying majorly on human language. It supports the connectionist perspective that language skills are the behavioural manifestation of the internal mental representations in the human brain that emerge from the interaction with the environment. Language and behaviour are yielded from simple learning rules that operate at a neural level, in lieu of hinging on pre-trained knowledge from large corpora. The design of the neuron model heeds computational efficacy over biological fidelity. An encyclopaedic description can be retrieved from the original work (Golosio et al., 2015).

Nearly all theoretical views acknowledge the undeniable existence of a long-term memory in the human brain (Chai et al., 2018). This holds facts and events, comprising the declarative knowledge. In analogy, the long-term memory (LTM) of the ANNABELL cognitive architecture holds information in the form of declarative phrases. The LTM storage includes a subnetwork for phrase memorisation and another for retrieving the memorised phrases, in the process of decision making (Figure 1). Albeit there is no complete distinction between short-term memory and the working memory, an oversimplified dissimilarity is that the working memory also includes information manipulation for complex cognitive utility, thus is responsible for behaviour (Cowan, 2014). There is a significant organisational analogy between the WM and ANNABELL. ANNABELL takes its inspiration from two revolutionary models of the working memory model, the Baddeley & Hitch model (Baddeley & Hitch, 1974; Baddeley, 1992) and the Cowan’s model (Cowan, 1995; Cowan, 2008). The former recognises three main components of the WM: the phonological loop (verbal working memory), the visuospatial sketchpad (visual-spatial working memory) and a controller which oversees manipulation between the former two subcomponents, called the central executive. These components can also be in part observed in ANNABELL. In the latter perspective, that of Cowan, the working memory is but a subset of an activated long-term memory, which is, in turn, an intermediate subset (short-term storage) of the long-term memory. It holds domain-specific activated elements (phonological, visual) bound in integrated units of information with chronological sequence by the focus of attention, a conjecture that is further supported by Baddeley (2010). This view approximates the methodology proposed in this thesis, by which a model like
ANNABELL lacking an explicit visuo-spatial sketchpad structure, can elaborate multimodal sensory information from the verbal, visual and motor domains (Chapter 4).

Moreover, there is a substantial procedural analogy amid the cognitive architecture and the working memory. The WM manipulates information of various nature to create a complete coherent thought, starting from modest capacities like recalling a sequence of digits in order, to doing complex arithmetic in the head (Cowan, 2014). These mechanisms in ANNABELL are modelled as elementary operations (phrase indexing, extracting words from and mapping them to buffers between the different subnetworks, retrieving a memorised phrase from the long-term memory, storing the working phrase as a goal chunk, etc.) that lead to task-oriented decision-making. These operations on word groups, phrase buffers and other subnetworks, called mental
actions, are triggered by action neurons and are combined in a sequence of mental actions, in five phases or modes of operation: the acquisition, association, exploration, exploitation and reward phase. The fundamental learning mechanisms and synaptic gating, which approximate ANNABELL with the principles of synaptic plasticity in biological neural networks and underpin its abstraction capabilities, are extensively described in the Appendix A.

Further justification of the framework

Two major approaches to language grounding in robotics include probabilistic modelling and the use of neural networks (Ogata et al., pp. 170, 2022).

Probabilistic models consider the relationship between language and other (non-linguistic) modalities as probabilistic relationships. The benefit of such representation is a high intelligibility. Each node in a graph model represents a meaningful element. This allows understanding the kind of inferences that are performed by the model. However, as a pitfall, probabilistic models cannot deal with long-term dependencies, i.e., problems whose outputs depend on inputs represented at an earlier time.

Neural networks represent the relationship between language and behaviour in the internal dynamics of the network. Most language neural networks include recurrent connections and gating mechanisms. This allows such models to learn long-term dependencies (as opposed to probabilistic models) without a priori knowledge. On the pitfall, their inferences are almost impossible to understand given that the representations in the hidden layers are distributed (black box), although some efforts to visualising their internal behaviour have been introduced in the literature.

Hence, the current methods of language grounding fail to provide high intelligibility whilst dealing with temporal structures. On the contrary, the framework leveraged in this thesis combines the advantages of both methods: a) as a large-scale neural network, it benefits from the characteristics of neural networks (e.g., learning long-term dependencies) and b) aids explainability of its outputs. Although the low-level behaviour of the model (internal activation state and connection weights) remains a “black box”, the training of the framework based on the mental action sequence allows to interpret and validate the creation of an output. The execution of the mental actions follows an explainable manner: it allows to follow the elementary operations computed by the framework at sentence level and the type of reasoning performed to achieve an output (see The mental action sequence, Appendix A and examples from the learning sessions, Appendix B).

Throughout this thesis, the term learning framework extensively remarked in the methodology chapters (3-5), refers to the ANNABELL cognitive architecture of Golosio et al., (2015) described here. The necessary documentation, the open-source code, and the user guide to interact with the framework can be retrieved at https://github.com/golosio/annabell/. This contains fundamental information about the original model. In this thesis, each chapter offers substantial enhancement to the procedural and memory retrieval mechanisms of the existing architecture, to accommodate the conducted research and its novel proposals described in each passage. These are appropriately documented and discussed in each chapter, respectively, and in Appendix A and Appendix B.
Chapter 3

Modelling Multiple Language Learning

3.1 Context

\textit{Per se}, the study described in this chapter was conceived to bring forth potential assumptions on the impact of multilingualism on the referent learning framework, in the multilingual acquisition and discerning competence, for instance, if the \textit{central executive} can oversee interference of the parallelly competing languages. Built-in the larger-scope research of the thesis, the chapter envisions the imprint of multilingual cognition in high-level processes (e.g., cognitive control), for example how different language inputs could alter high-level phenomena (abstraction, inference, categorisation) during decision-making that may result in significant consequences for cognition and the intelligent behaviour of robots.

The chapter presents a methodology for acquiring multiple human languages, which is extensively validated in a series of learning experiments with the selected learning framework. The \textit{first set} of experiments (part 1) is conducted with a rather large quadrilingual corpus that describes an assumed social environment. In the first experiment of this set, the phrases of each language are learned and tested separately (\textit{monolingual acquisition}). In the second experiment, languages are trained and tested jointly (\textit{polylingual acquisition}), and the acquisition degree of skill is contrasted with the monolingual competence. The third trial investigates the gradual acquisition of languages and compares the multilingual competence in each stage (\textit{mono-, bi-, tri- and quadrilingual acquisition}). The \textit{second set} of learning experiments (part 2) is conducted with a smaller dataset in the constructive format of a narrative passage (pre-school story) to validate the competence for coherent dialoguing. The experimental set is divided into two stages: in experiment 1, the story is trained and tested in each language separately (\textit{mono}), whereas in experiment 2, the story is learned in multiple languages jointly and the content is tested at random language (\textit{poly}). The results of the two experiments are compared contra one another. The methodology is first applied in a computer-based machine interface and next is suitably modelled in a robotic architecture using a smaller corpus from the first dataset (part 3). It aims to explore the potential of such multilingual modelling for human-robot bilingual interactions and their applicability as social companions.

The research described in this chapter is published in \textit{Giorgi et al., 2020a} and \textit{Giorgi et al., 2020b}.

3.2 Prologue

3.2.1 The Hierarchy of Knowledge Representation

\textbf{Anderson & Krathwohl, (2001)} have proposed a hierarchy of the three major types of knowledge (\textbf{Figure 2}). This hierarchy is applied in the methodology described here, to train the learning framework the skill of usage-driven language acquisition. The \textit{declarative knowledge} consists of a series of affirmative statements and natural cues on how to use them in conversation. The \textit{procedural knowledge} for language tasks is obtained by training \textit{the proper way} to answer simple questions, using preschool-level phrases as examples during communicative interactions with the
human interlocutor. This learning approach complies with the Natural Approach of Language Learning (Krashen, 1989), which posits that grammar rules are not requisite when first acquiring the language. Rather, continuous exposure to the language and how it is properly wielded in everyday situations leads to the spontaneous emergence of verbal expression. Similarly, the methodology proposed here focuses not on learning the grammatical constructions of the languages; instead, syntactical, and semantical soundness are yielded from how language is used in contextual verbal exchanges with the human. The generalisation skill emerges by following the same line of reasoning in sentence production and recalling the learned experiences in close but unseen endeavours (metacognitive knowledge).

![Declarative Knowledge](image)

Declarative Knowledge
Factual knowledge in the form of true affirmations

![Procedural Knowledge](image)

Procedural Knowledge
Imperative knowledge or the skill of performing some task

![Metacognitive Knowledge](image)

Metacognitive Knowledge
The ability to use or relate past experiences in similar unseen tasks

Figure 2 The three major types of knowledge as defined by Anderson & Krathwohl. (2001).

### 3.2.2 The Selected Human Languages

The languages considered in this work have distinct complexity, where complexity generally is defined as the number and diversity of elements, along with the intricacy of their inter-relation structural (Kortmann, 2012). Here it is regarded as:

1. **Syntagmatic/lexical complexity**: word length, composite words structures, prepositions, different degrees of deixis.
2. **Morphological complexity** (word formation), e.g., fully irregular plurals.
3. **Organisational complexity**: component arrangement (e.g., adjectival order) and the word order within a sentence.
4. **Semantic complexity**: the meaning of the words might change in the way they are arranged in the sentences.

Although the languages belong to the same (but distinct branches of the) Indo-European family, a language from a different family does not affect the methodology and the overall learning mechanisms. It would only require careful consideration of its complexity (of all aspects 1-4 above) when constructing the learning corpus. Here, the focus is not to study the underlying structures of the languages (complexity or organisation), but rather to investigate how this affects the performance and skill of multiple language acquisition. The linguistic competence will depend on fine training data. Building adequate knowledge by experience via meaningful training samples requires proper identification of the semantic expectations that underlie all language-related tasks. All elected languages cover complex linguistic patterns that intercorrelate further among one another and spiral the languages’ intricacy. This is explained below:

The Greek language features three gender types (masculine, feminine, neuter) and four noun cases (nominative, genitive, accusative, vocative), 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 forms compounds flexibly and tends to be periphrastic (usually for future tense), which can be an arduous task as the meaning depends on the different number of words or word order in a phrase. However, the monolectic compound-constructing
capability of the Greek language, i.e., the property that single compound words can convey the meaning expressed by an entire sentence or paragraph, is habitually advantageous.

The Italian language possesses similar features in terms of gender, number and case for nouns, verbs (high degree of inflection) and adjectives. Conjugation (usually three patterns) is affected by person, tense, number, mood, aspect and occasionally gender. In addition, 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 necessary for clarity or to add desirable emphasis or contrast. Direct and indirect object pronouns (that receive direct or indirect actions), cannot stand without a verb.

In the Albanian language, 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. There are 6 noun cases (nominative, accusative, genitive, dative, ablative, vocative), introducing a change in the word structure that may be difficult to address. The plural is generally irregular.

Tackling the linguistic perplexities of each language in the learning stage requires adequate construction of the sentence structure and context, which affects the dataset configuration in the number and type of meaningful training examples for proper acquisition and how the mental action sequence is trained to build a valid output. Proper disambiguation of the languages is instrumental when languages are acquired jointly. Most of the linguistic properties discussed in this section are accounted for when building the learning corpus in this work. When modelling this learning, the learning framework must be equally exposed to ample quantities of comprehensible input required for each language that results in the accurate acquisition of that language (e.g., irregular plurals, genders...), even when this is not required in any of the other languages (e.g., English).

3.3 Multilingual Modelling Methodology: Part 1

The methodology is drawn using the learning framework described in section 2.3.4.1. Contra to its originally demonstrated linguistic functionalities, here this framework is exhaustively trialled to achieve the competence for suitable cognitive elaboration of multiple languages concurrently, a type of multilingual modelling that is poorly (if at all) explored in the existing literature (2.2.3). The methodology proposed here assumes that there is a level of language processing that involves the flow of information among the working memory buffers, which is language independent. As such, the method grants a general solution for the joint learning of parallelly competing languages, where the global organisation and the procedural mechanisms used in verbal elaboration remain intact (i.e., the ability to control the flow of information among memory components and buffers). The procedural knowledge is, however, dependent on the language that influences the number and type of mental action sequences that must be learned to acquire proficiency in each language, when learning them jointly. The languages are simultaneously fed without distinction or explicit indication (information is handled as if it were of the same language). While word-concurrence is not relevant in this framework (as words are represented as orthogonal vectors), phrase structural similarity can significantly affect the proper construction and retrieval of the action sequences, which can lead to competing information retrieval from memory and intermediate buffers and, thus, challenge the procedural mechanisms.

3.3.1 Dataset 1: A fictional multilingual social environment

The dataset is devoted to the thematic group People, originally described in Golosio et al., (2015). It is in part inspired by the Language Development Survey (LDS) work of Rescorla and colleagues.
(Rescorla & Alley, 2001; Rescorla & Achenbach, 2002), which provides valuable insight on the number and types of words or word combinations known and used spontaneously by toddlers. The LDS is used here to construct a systematic dataset. The focus of the work is language acquisition and no classical tasks studied in natural language processing (NLP), ergo the dataset is drawn using vocabulary that occurs naturally in the process of language development, over standard NLP datasets. The generated corpus is suitable for an extensive quantitative evaluation on the multiple language acquisition by the cognitive architecture, using a simple lexicon at a developmental level of complexity. The sentences are appropriate for conversation with a pre-school child, capable but no expert in delivering information in conventional ways.

**Declarative or explicit knowledge**

The declarative knowledge of the learning framework is fed in the form of declarative/factual affirmations that are used to verbally describe the situated social environment of a fictional 4-year-old girl called Annabell. The supposed environment includes twenty people and nine potential relationships (mother, father, friend, ...) with the main character. These verbal affirmations include descriptive phrases such as “you have a mother”, “your sister has an umbrella”, “you like cats”, “your father’s name is”. This lexicon is stored in long-term memory and the learning framework is queried with simple questions in its conformity.

Some declarative (prescriptive/how-to) sentences guide the natural way to correct usage of the language, without teaching or imposing specific grammatical rules; for instance, *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 phrases aid sentence production and are self-retrieved from the long-term memory during question-solving. For instance, when queried on age comparison tasks, the learning framework recalls how to compare numbers. From the cognitive physiology angle, this resembles the type of knowledge acquired from experience and transferred to new contexts (e.g., naturally using pronouns to refer to people and their possessions, without literal awareness of what a pronoun is).

The declarative phrases are translated in all four languages and jointly comprise the knowledge in explicit form. The translated phrases preserve content but are placed in an appropriate language-imposed context. Otherwise speaking, they are structured in a way that addresses the different morphosyntax or introduces where needed distinct uses of plurals, genders, noun cases, verb conjugations or “pro-drop” forms that are naturally specific to each language.

**Procedural or implicit knowledge**

The human-machine conversation is shaped in the form of question-answering. The training methodology is inspired by parent-child verbal interactions similar to how parents use simple communicative examples to query the child about the world as they know it, rather than teaching the grammar constructions or specific linguistic forms. However, here it is not modelled a real child’s talk that emerges from real-world interactions, but rather simple meaningful phrases that can be known and elaborated by a four-year-old. Note that training here does not refer to the kind of training employed in much of the machine learning literature. Instead, it refers to the series of interface commands executed by the human user, which are converted into elementary operations computed by the framework at sentence level (see Appendix B).

Training is organised in two stages. In the first stage, the learning framework receives through the computer-based interface and stores in memory a set of declarative sentences (explicit knowledge) that describe the social environment. In the subsequent training stage, the human user guides the framework to learn to answer a set of sample questions related to the explicit knowledge, i.e., build
the mental action sequence to obtain valid answers on the requested phrases (implicit knowledge). A sample construction of the mental action sequence is illustrated in Algorithm 1 and its low-level (neural level) implementation is described in Appendix B.

**Algorithm 1** High-level pseudocode of a sample construction of the mental action sequence for a specific input.

```
Algorithm 1 EXAMPLE OF A MENTAL ACTION SEQUENCE
1: Input: how old is your sister
2: Output: she is seven years old
3: EXTRACT wordGroup your sister
4: RETRIEVE memorisedPhrase Susan is your sister
5: if memorisedPhrase then
6: SET workingPhrase Susan is your sister
7: EXTRACT wordGroup Susan
8: RETRIEVE memorisedPhrase the personal pronoun for Susan is she
9: if memorisedPhrase then
10: SET workingPhrase the personal pronoun for Susan is she
11: EXTRACT wordGroup she
12: REWARD partialREWARD > send first word to the output
13: end if
14: SET workingPhrase Susan is your sister
15: EXTRACT wordGroup Susan
16: RETRIEVE memorisedPhrase Susan is seven years old
17: if memorisedPhrase then
18: EXTRACT wordGroup is seven years old
19: REWARD fullREWARD > send full phrase to the output
20: end if
21: end if
22: MEMORISE stateActionAssociation
```

The interrogative phrases used in the dataset are also inspired by the work of Rescorla and colleagues. They are divided into a training set and a test set. There are 46 distinct types of phrases that are used to train the linguistic skills of:

1. using pronouns to refer to people and objects;
2. answering polar, multiple-choice and wh-questions, e.g., what is your sister’s name;
3. age comparison tasks, counting and comparing numbers, e.g., who is older/younger;
4. telling its own likes and dislikes, e.g., do you like (to) …
5. recognising other people’s likes and dislikes (and the types of possible relations between them and the main character), e.g., does your father like to drive;
6. recognising different professions of the involved persons, e.g., Matt is a driver

For each question Q, there are at least 4 phrases similar in structure and context, but using different open-class words. One (or at least one) of the sentences is used for learning, while (at least three of) the remaining of each type are used to probe the generalisation competence (Table 1).

**Table 1** The multilingual corpus. The number of declarative sentences used to set/describe the social environment (column 2), the number of interrogative sentences used for training (column 3) and the number of questions used in the test stage (column 4), in each language (column 1).

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Declarative phrases</th>
<th>Training set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>308</td>
<td>89</td>
<td>292</td>
</tr>
<tr>
<td>Italian</td>
<td>319</td>
<td>89</td>
<td>292</td>
</tr>
<tr>
<td>Greek</td>
<td>355</td>
<td>89</td>
<td>292</td>
</tr>
<tr>
<td>Albanian</td>
<td>364</td>
<td>89</td>
<td>292</td>
</tr>
</tbody>
</table>

The interrogative phrases of the training and test sets are translated in each language, with the same criterion applied to the declarative statements. To achieve a fine acquisition of each language in accordance with their phonological, morphological, and syntactic structure, it is often needed to increase the number of training samples that enable accurate generalisation, for instance on gender, plural, case, and conjugation. Let us consider a learning case with the use of the pronoun which when replying to a question, e.g., which cousin. In English, which is gender-neutral; however, in Greek and Albanian, the pronoun changes form with the gender of the referenced noun (e.g.,
poia/poios in Greek). For an accurate evaluation of the degree of skill in each language, the number of training and test samples per language was kept equal, i.e., equal exposure (columns 3 & 4, Table 1). The size comparison between the original dataset and the compound multilingual corpus is illustrated in Figure 3.

![Figure 3 The original (English) and polylingual corpora. The language datasets are normalised to involve equal lexicons and learning samples.](image)

### 3.3.2 Experimental Validation

The experiment aims to assess the degree of skill of the learning framework in joint multiple language acquisition. It is important to emphasise the concept “acquisition” used here 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 (Kortmann, 2012). Language acquisition best describes the type of training performed in ANNABELL that is in the form of communicative activities, in which the system experiences how language is used and reproduces it closely to communicate back.

To demonstrate whether the simultaneous acquisition of all languages affects language processing and production functionalities, the experimental setup is divided into two fragments:

1. The languages are initially trained separately, and the performance is measured at the end of each individual test (monolingual acquisition).
2. The languages are trained jointly, and the performance is measured at the end of the cumulative test (polylingual acquisition).

The two parts are carried out independent of one another to compare the behaviour of the learning framework when it acquires a language alone or jointly with others and to study the potential interference in the processing of the information given in different languages. The accuracy criterion was defined as the fraction of valid answers over the total number of requested phrases, where a valid answer is one that is both syntactically and semantically correct and appropriate for the conversation and language.

#### 3.3.2.1 Monolingual Acquisition

**Setup**

The quantitative performance evaluation was performed through a *k-round cross-validation* (CV) technique. The cross-validation is organised in four sessions (rounds). The rounds are executed
separately, starting from a clean slate and the final performance is then averaged on all four rounds. This is repeated for each language disjointly.

Each round follows three stages of execution: (a) the learning framework acquires the declarative statements that describe verbally its fictional social environment and assumed past experience; (b) it is appropriately trained to answer a set of distinct questions related to the declarative statements, using one (or at least one) learning example per each question type; (c) it is probed with a previously unused set of questions similar in structure with those learned in stage (b), but involving different open-class words. To obtain the rounds independently the learning sample differs in each round. There are at least four different interrogative phrases of the same type. One of them is randomly extracted for the training set and the remaining are used in the test, with the constraint that the same training question should not be used in different rounds. This allows assessing the system behaviour four times independently, by varying the training and test sets.

The order in which the questions are learned and probed is randomised. An important feature of the procedural mechanisms of the learning framework is the ability to build an output preferentially by executing mental operations during decision-making. When asked to solve a task, i.e., answer to a question, it needs to retrieve, extract, and compare the phrase(s) from the long-term memory that are most appropriate for the type of input query. The framework does not exploit correlations among words or next word predictions, because words are represented in the architecture as orthogonal vectors. Instead, the construction and exploitation of the mental actions in information elaboration resembles how the human working memory binds together pieces of information to create a complete and coherent thought. This prompts the generalisation capability of the learning framework in language acquisition and the resemblance with the natural way that humans process verbal information in the brain.

**Results**

The results of the monolingual acquisition are illustrated in Figure 4. Each of the vertical bars represents the percent accuracy averaged over the four rounds of the cross-validation for that language. The standard deviation is also shown. The exact numbers of correct output phrases produced in each round and for each language are given in Table 2.

![Figure 4](image_url) *The quotient of valid answers produced by the learning framework in each language. The accuracy values are averaged over the four rounds of the cross-validation and the dispersion is indicated by the standard deviation.*

<table>
<thead>
<tr>
<th>Monolingual acquisition</th>
<th>Accuracy of monolingual acquisition averaged over the 4-rounds of cross-validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>89.38%</td>
</tr>
<tr>
<td>Greek</td>
<td>84.42%</td>
</tr>
<tr>
<td>Italian</td>
<td>86.04%</td>
</tr>
<tr>
<td>Albanian</td>
<td>86.30%</td>
</tr>
</tbody>
</table>

*Table 2* *The fraction of valid answers produced in each language in each round of the cross-validation. A valid answer must be both syntactically and semantically sound.*
For a deeper understanding of what type of questions were poorly elaborated by the learning framework, it was investigated how errors were distributed across the corpus. This is illustrated in Figure 5 for rounds one through four. Each small-scale chart includes all languages side by side for comparison, where the vertical bars correspond to the number of invalid output phrases for the type of requested input (type of question Q). Some input question types with error-free output phrases across languages are not reported.

There is a uniform error distribution across the test set in each round of each language. Peak errors occurred for high-complexity tasks. In question types 34 (do you like ...) and 35 (does <person> like ...), the Greek and Italian languages would only generate a valid answer for slightly above half the total queried phrases. Instead, the Albanian language rendered relatively favourably in Q35, as dexterously as or outperforming English in some rounds (3 and 4). Although the intricacy of the language is higher, so is the regularity of sentence structure of type Q35, which is not the case for Q34, resulting in destitute accuracy.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
<th>Round 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>89.04%</td>
<td>90.41%</td>
<td>90.41%</td>
<td>87.67%</td>
</tr>
<tr>
<td>Greek</td>
<td>81.16%</td>
<td>86.99%</td>
<td>84.59%</td>
<td>84.93%</td>
</tr>
<tr>
<td>Italian</td>
<td>83.19%</td>
<td>88.70%</td>
<td>88.70%</td>
<td>83.56%</td>
</tr>
<tr>
<td>Albanian</td>
<td>85.62%</td>
<td>85.62%</td>
<td>86.30%</td>
<td>87.67%</td>
</tr>
</tbody>
</table>
Submitted Thesis: *Human Language as a tool for Conceptual Development in Cognitive Robotics*

Figure 5  **Distribution of errors across question types in each disjoint language corpora and all four rounds of the CV.** All languages have a uniform error distribution spanned on all rounds, with steady peak values in complex tasks 34, 35, 41, and 42 and random peaks elsewhere. Error-free tasks are not shown; they recurrently relate to phrases of simple structure, of equal length and word order, and unvarying grammar.

The higher error distribution for some tasks was expected and can be justified with a) the higher perplexity of languages, which b) the limited number of learning samples is insufficient to address exhaustively. For instance:

1. Greek, Italian and Albanian depend heavily on gender and number for articles and verb conjugations. In English there is no significant distinction between the following phrases *you like cats*, *you like chocolate*, *your sister likes dolls*, *your father likes TV*. Shall we investigate the structure of the respective phrases in Italian: *a te piacciono i gatti*, *a te piace il cioccolato*, *a tua sorella piacciono le bambole*, *a tuo padre piace la TV*.

![Diagram of language structures](image-url)
Figure 6 Sample illustration of grammar & morphosyntax dissimilarities between English (up) and Italian (down).

The above phrases in English introduce only one change in person (second/third). The respective Italian phrases include two distinct verb conjugations, which vary with the number of the referent noun object, two gendered pronouns that vary with the referent noun subject and four articles that agree in gender and number with the referent object. The example illustrated in Figure 6 remains true for Greek and Albanian. It is reasonable to infer that higher exposure to learning samples might lead to enhanced generalisation competence.

2. A property of the Greek language is that masculine nouns 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. ὁ ἄντρα (man) would change with the case: nominative: ο ἄντρας, accusative: τὸν ἄντρα, vocative: ἄντρα. Should the selected training input involve a masculine noun, learning must account for the form change, i.e., if the noun in the question is expressed in one case and the output sentence requires using that word in another case (a change in form, reflected in a further step of the mental action sequence), the learning framework cannot generalise on feminine nouns, which do not impose the form transition. The vice versa is also true. Consequently, not only is the size of the training samples central, but the learning sample itself overwhelms the linguistic competencies of the learning framework.

3. A habitual error that occurred for languages besides English stemmed from their irregular plural-formation. The working memory of the framework is more heavily tasked when using both forms (plural and singular) in the same context, which is reflected in the performance decline.

3.3.2.2 Polylingual Acquisition

Setup

In the monolingual cross-validation, it was demonstrated that the learning framework acquires any human language competently with no cost on architectural changes and the general neural processes that underpin how the system elaborates information. The questions that the polylingual cumulative cross-validation intends to address are:

1. To what extent or skill can the system acquire multiple languages simultaneously, disambiguate and identify the (dis)similarities of the languages?
2. Can it self-infer the most appropriate response to a random question in an unanticipated language, without confusion and/or without altering the linguistic functionalities of either language?

3. Does learning languages jointly affect the ability to acquire a language satisfactorily and comparably to its monolingual acquisition?

The term cumulative used here refers to training all languages datasets jointly before probing them in random language. A four-round open-ended cumulative cross-validation was carried out to measure the acquisition performance. This relatively long developmental process is sustained by the large-scale of the network and the ability of the learning framework to perform real-time communication. The rounds are built similarly as explained in section 3.3.2.1, however here each round includes the lexicon of all languages (e.g., the disjoint 1st rounds of English, Greek, Italian and Albanian supply the 1st round of the cumulative cross-validation). The rounds were executed independently, and the total accuracy was averaged at the end of all rounds. The size of the aggregate polylingual corpus is illustrated in Figure 3. The lexicon covered 900 different new words from all languages, with an average of 6.263 words per sentence.

Results

The obtained results in Figure 7 indicate a slight decline in the overall accuracy compared to the monolingual acquisition. This accuracy corresponds to the calculation described by equation (1).

\[
\text{accuracy} = \frac{\text{number of valid output phrases in one language}}{\text{total number of requested input in that language}}
\] (1)

Another formulation of the accuracy criterion is how many valid phrases does the multilingual framework generate in each language out of the total times it is probed in that language. Thus, it is useful to compare with the ratio of correct output produced when languages are acquired alone.

Table 3 The fraction of valid answers produced in each language in each round of the cumulative cross-validation. A valid answer must be both syntactically and semantically sound.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
<th>Round 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>79.11%</td>
<td>85.62%</td>
<td>86.64%</td>
<td>83.46%</td>
</tr>
<tr>
<td>Greek</td>
<td>76.08%</td>
<td>83.25%</td>
<td>76.71%</td>
<td>77.74%</td>
</tr>
<tr>
<td>Italian</td>
<td>77.74%</td>
<td>74.66%</td>
<td>85.27%</td>
<td>74.32%</td>
</tr>
<tr>
<td>Albanian</td>
<td>77.74%</td>
<td>78.39%</td>
<td>77.74%</td>
<td>78.05%</td>
</tr>
</tbody>
</table>

The performance loss regarding the monolingual acquisition was comparable for all languages: Greek 5.91%, English 6.05%, Italian 8.04%, and Albanian language 8.33%. This decline was expected because the size of the aggregate dataset is four times larger than each of the monolingual datasets, thus affecting the number of learnable interconnections that are created and loaded from the framework. The links saturate much faster, which may cause a mild interference in the processing of the acquired information and can be mitigated by increasing the number of input examples in the training stage and/or increasing the number of neurons and connections. What is more, information becomes less homogenous and linguistic structures of the different languages might correlate significantly.
Chapter 3: Modelling Multiple Language Learning

Figure 7 The quotient of valid answers produced in each language when trained jointly. The values are averaged over the four rounds of the cumulative CV and the dispersion is indicated by the standard deviation. The contrast with the monolingual acquisition is displayed to assess at what degree of skill (in terms of accuracy) does the learning framework acquire the languages jointly over each alone.

*Embryonic evidence of code-mixing*

The results revealed cases in which the multilingual framework would produce hybrid-language answers, where the root meaning for the requested conversation changed slightly or barely. Below are illustrated three meaningful precedents.

i. the output phrase does not answer the question, but is syntactically correct, expressed in a mixed language;

**English-Italian code-mixing**

*Human:* Chi è Letizia?
*Annabell:* lei is a girl

**English translation**

*Human:* Who is Letizia?
*Annabell:* she is a girl

ii. the answer is syntactically & semantically correct and appropriate for the conversation, but in incompatible language;

**English-Italian code-mixing**

*Human:* Quanti anni ha Susan?
*Annabell:* she is seven years old

**English translation**

*Human:* How old is Susan?
*Annabell:* she is seven years old

iii. the answer is semantically yet not syntactically correct e.g., two synonymous words from different languages are merged within the produced phrase, which yet fit the context;

**Greek-Albanian code-mixing**

*Human:* A ka Susan një biçikletë?
*Annabell:* Podildato nuk ka një biçikletë

**English translation**

*Human:* Does Susan have a bike?
*Annabell:* Bike does not have a bike

Whilst *code-switching* or *code-mixing* is regarded in the literature as a natural part of bi-multilingualism (*Krashen, 1989*) there is no convincing evidence that this is the case in the learning framework, hence these cases were classified as errors. However, it is particularly interesting to note that there is certain relevance to the above scenarios, at either the grammatical (syntactic or morphological) level or the word (semantic) level. It might indicate that it discerns the context of the question or some correlation between the languages and switches or “borrows”
words to construct a valid sentence. In turn, it may suggest that the neural representations of the languages are activated simultaneously, resulting in interference from the competing languages.

### 3.3.2.3 Piecemeal Cumulative Learning

**Setup**

Seeking to understand deeper the potential causes of the accuracy loss in the cumulative training, a piecemeal cumulative training was adopted. Given that the results of cumulative cross-validation (polylingual acquisition) can be affected considerably by the size of the corpus over the number of neurons in the subcomponents and of their inter-connections, training languages gradually grants a finer qualitative evaluation of the multilingual framework’s linguistic competencies.

The languages were arranged in 4 datasets of augmenting size: monolingual, bilingual, trilingual, and quadrilingual. Each consecutive dataset comprises the languages of the previous lower-level dataset. The datasets were trained and tested disjointly. The results of this training are given in Figure 8, on round 1 of the cross-validation. As the number of sequential combinations with two, three and four languages are large, the proposed scenario, here assumed significant, first comprised the languages of higher intricacy that unveiled lower accuracies in the cumulative CV: Albanian (mono), Albanian-Italian (bilingual), Albanian-Italian-Greek (trilingual), all (quadrilingual).

**Results**

The results of piecemeal training unveiled finer results for complex languages, with little or no variance of accuracy between learning two or three languages (Figure 8). 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 the number of neurons and interconnections affect storing and processing significantly large corpora. The learning framework could generalise satisfactorily from one or a few instances of training samples, on an eclectic lexicon that included nouns, verbs, adjectives, and other open-class words. This advocates that acquiring a language jointly with others does not significantly affect its ability to acquire the language as adroitly as acquiring it alone.

![Figure 8](image-url) The elaboration competence of the model in augmenting multilingual lexicons. The proposed scenario assumed as significant first accounted languages of higher intricacy, gradually scaling up the dataset with more languages, one at a time. The end performance of the quadrilingual dataset was faithful to the results of round 1 in the cumulative cross-validation (column 1, Table 3), suggesting that the chronology in which languages are learned is imperceptible.
3.4 Multilingual Modelling Methodology: Part 2

Machine comprehension of texts and the ability to answer context questions is an open problem in AI and human-machine interaction. From a psychological perspective, the understanding of narratives requires assessing what people recall from the story and their response to probe words (Wyer, 2014).

This section describes an artifice with the learning framework, which seeks to explore its potentiality to elaborate a narrative passage to support coherent dialoguing. This is not a traditional machine reading comprehension of a text e.g., from public datasets or benchmarks, but limited to the level of child comprehension and pragmatic in a multilingual scene.

3.4.1 Dataset 2: A multilingual narrative passage

The narration is extracted from the book “My first jungle story” (Watson, 2019) of preschool literature. The content is translated in the 4 languages of choice (section 3.2.2) to construct the datasets. The plot 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. Put simply, these are questions that require longer comprehensive formulations over bipolar (yes/no) answers.

A major difference with the explicit knowledge in dataset 1 (3.3.1) is that the declarative phrases used here to describe each of the animals and their story communicate a continuous coherent and cohesive meaning throughout the narration. Each question follows the preceding and can only be fathomed in the currently running context. After learning about all the animals, the learning framework is trained to answer three consecutive questions with logical and chronological continuum. The interrogative phrases are inspired and/or originally extracted from the story itself.

i. what did Leo learn about the <animal>, where <animal> is any of the 9 characters of the story;
ii. can it <verb>, where it indirectly refers to the animal of the previous question and <verb> is an attribute of that animal, e.g., can it fly/swim/dig;
iii. what can it do, where the pronoun it refers to the same animal from questions i-ii;

Although text comprehension also presumes making prepositions that derive from the story, which might further relate to previous knowledge stored in long-term memory and is often not found within the text itself (Wyer, 2014), this is not a target of the experiment. Rather, the focus here is to explore the following three aspects of text assessment:

1. The competence to answer open-ended context-dependent questions meaningfully and describe the animals using manifold full-length sentences.
2. The ability to dialogue coherently and cohesively with the human, while relating questions with one another and within the context, and tracking the referent (it) of the conversation.

Compellingly, it is investigated if the learning framework can self-infer the referent animal in the conversation (an attribute of the working memory) when it is probed in the hereunder cases: (a) the human refers to the animal as “it”, but offers an attribute as a cue (e.g., fly, swim, dig...); (b) the human poses an ambiguous question (what can it do), with no explicit hint on the animal, i.e., any animal can be an answer candidate, but a single fits the coherent context. An intuitive question to ask: How can it conjecture the referent?
3. The generalisation capability of the learning framework on a previously unused lexicon from the story in any language (*monolingual acquisition*) and/or in unanticipated language (*polylingual acquisition*).

### 3.4.2 Experimental Validation

The 4-fold cross-validations (mono/polylingual) are performed faithfully to what was explained in sections 3.3.2.1 and 3.3.2.2. The learning sets for each round are constructed using a new animal as an input sample (to train questions i-iii) while the remaining nine animals are used in the test. The rounds are executed independently. **Figure 9** illustrates the comparative accuracies unveiled in the test during monolingual and polylingual acquisition.

There was some evidence of resemblance with code-switching in the polylingual acquisition. For instance, due to the identical spelling of the word *Hippo* in Italian and Greek (*Ippopotamo*), the multilingual framework misuses the pronoun *it* but builds a valid answer at the semantic and syntactic levels.

**Figure 9** Context-oriented mono and polylingual acquisition of a narrative story. The ability to learn the story in a mono language was compared against learning it parallelly in multiple languages. The accuracy is averaged over the four rounds of cross-validation and its fluctuation is indicated by the standard deviation. In the multilingual scenario, the learning framework was probed randomly in any language, attending to the criterion that related questions must be of the same language, whereas the language can vary with the queried animal.

<table>
<thead>
<tr>
<th>Language</th>
<th>Monolingual</th>
<th>Polylingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>90.74%</td>
<td>87.04%</td>
</tr>
<tr>
<td>Greek</td>
<td>86.11%</td>
<td>92.59%</td>
</tr>
<tr>
<td>Italian</td>
<td>88.89%</td>
<td>81.50%</td>
</tr>
<tr>
<td>Albanian</td>
<td>86.11%</td>
<td>80.56%</td>
</tr>
</tbody>
</table>

**Human:** ti emathe o Leo gia ton Ippopotamo (= what did Leo learn about the Hippo)
**Model:** esso eiche ena megalto stroggylo soma (= it had a large round body)

In other cases, the model would fail one of the queries in the chain i-iii but continue to attend to the dialogue for successive questions. This is sustained from the ability to retain the task goal and retrieve it at a later stage of decision-making even if some intermediate output is not built correctly.

<table>
<thead>
<tr>
<th>Human:</th>
<th>cosa ha imparato Leo della Antilope?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model:</td>
<td><em>void</em></td>
</tr>
<tr>
<td>Human:</td>
<td>essa puo ballare la samba?</td>
</tr>
<tr>
<td>Model:</td>
<td>No essa non puo ballare la samba</td>
</tr>
<tr>
<td>Human:</td>
<td>cosa puo fare?</td>
</tr>
<tr>
<td>Model:</td>
<td>essa puo saltare in alto e sopra le rocce (= English translation)</td>
</tr>
<tr>
<td>Human:</td>
<td>what did Leo learn about the Antelope?</td>
</tr>
<tr>
<td>Model:</td>
<td><em>void</em></td>
</tr>
<tr>
<td>Human:</td>
<td>can it dance the samba?</td>
</tr>
<tr>
<td>Model:</td>
<td>No it can not dance the samba</td>
</tr>
<tr>
<td>Human:</td>
<td>what can it do?</td>
</tr>
<tr>
<td>Model:</td>
<td>it can jump high and leap over the rocks</td>
</tr>
</tbody>
</table>
The test results in Figure 9 demonstrated that the learning (multilingual) framework could track the correct referent across the context, describe animals meaningfully when probed with open-ended questions and generalise appropriately on each character. The property to dialogue in such a manner is attributed to its ability to store a sequence in the goal stack (typical in cognitive architectures and the working memory) when a mental action cannot be performed immediately. During decision making, it can recognise that one word in the current phrase is equal to a word of the phrase stored as a goal and place it in the appropriate niche to build the valid answer. Moreover, it could maintain its performance when learning the story in different semantic ambient, with an accuracy above 80% and comparable to the monolingual competence. The unwavering standard deviation suggests that there are negligible statistical fluctuations, which can derive from the small number of samples in the training and testing stage. Although the corpora are much smaller than the ones described in section 3.3.2, thus bringing about superior performance, it is worth noting that all languages are elaborated at a comparable degree of skill despite their intricacies. Most errors were related to code-switching likeness. An unforeseen phenomenon occurred in the Greek language, for which it was intuitively expected that the accuracy would decay in the multilingual context. It can be reasoned that given the peculiarity of the language, training jointly with other languages of similar complexity and morphosyntax strengthened the framework’s ability to generalise on Greek.

3.5 Human-Robot Bilingual Interaction: Part 3

Here, the multiple language learning modelled in the previous sessions is appropriately employed for real-time interactions with a social robot. The cumulative acquisition is narrowed down to two languages, English and Italian. The aim was to investigate the efficacy of a bilingual social robot to communicate in a real-world configuration using a pre-school level lexicon. In their work “Why Robots?”, Cabibihan et al., (2013) claim that robots can shape better interpersonal communication with human users and greatly assist in social domains. Here the desired intent is that the robot (a) automatically recognises the spoken human language; (b) disambiguates the languages in semantic and grammatical level with a high confidence score to build valid answers; (c) responds correctly in the obliged semantic context.

3.5.1 Robotic artefact

Q.bo One is a low-cost and ROS-based programmable robot (The Corpora, 2022). Its open-source software platform operates on Linux (Open Q.bo) and the core design integrates a Raspberry Pi with an Arduino 0 board. Q.bo is intended for artificial intelligence and social development robotics that is undergoing constant improvement.

The robot (Figure 10) lacks a humanoid form, with no limbs and therefore is more suitable for simple verbal interactions and recognition tasks, supported by its stereoscopic vision that allows face and object detection and tracking. This feature, along with the speech synthesis and natural language processing, enabled by three microphones and a sound card for sophisticated sound manipulations, makes the robot an adept social companion that strives to maintain eye contact with the humans and is able to listen to them talking from anywhere in the room.
Q.bo One robot is equipped with LEDs, which imitate mouth movement during speech production. Two digital servos control up-down and left-right movements of the head and capacitive sensors on the side let the robot react to touch. Though unable to perform body gestures, the robot can reinforce conversations with personable non-verbal attachment with the human via face tracking with head movement and eye contact. While robots like Nao/Pepper often used in social contexts are more appealing given their humanoid form and ability to “read” emotions, they require a lot of power only to stand up, resulting in their motors heating up quickly (Mubin et al., 2018), which is not justified for the type of verbal engagement described here. Though coupled with a low-cost robot such as Q.bo One, the interface artifice and learning method devised here can be employed for more sophisticated robots like iCub (IIT, 2022).

3.5.2 The Instantiated Robotic Model

3.5.2.1 Automatic Language Recognition and Speech Synthesis

The Google Cloud Speech APIs (Google, 2022) is used to set up speech recognition and synthesis of the robot. This selection is justified given the APIs functionality of detecting spoken language automatically, which allows multiple language recognition of audio transcripts and enables simultaneous bilingual interactions with the human. Google Speech-to-Text uses powerful neural network models to convert audio to written content. The API is lightweight and supports over 120 languages and dialects. It can transcribe real-time streaming or pre-recorded audio (stored locally or remotely). Google Text-to-Speech converts text into human-like speech in more than 30 languages and dialects, with 180+ humanlike voices, to deliver high-fidelity audio.

3.5.2.2 Yet Another Robotic Platform (YARP)

The communication platform used to bridge the learning framework (ANNABELL) with the Q.bo robot and the speech APIs is the Yet Another Robot Platform (YARP) interface (YARP, 2022).
YARP is an open-source software package for interconnecting actuators in robots. The framework includes a dedicated interface for coupling with YARP.

The YARP module is a transport neutral model of communication, and the data flow is decoupled from the details of the underlying networks and protocols in use, operating simultaneously, which grants a smooth evolution. The interface allows for several parallel TCP/IP connections (read/write ports) and encourages loose coupling of devices to bridge the different components of the final fully embodied robotic model displayed in Figure 11.

3.5.2.3 The Flow of Information in the Cognitive Robotic Model

On one end (left side), the robot handles Speech APIs and transcript exchanges over the duplex SSH connection. The YARP interface is responsible for passing information to/from the cognitive system, operating jointly at a remote computer on the other end (right-side).

To achieve higher speech recognition confidence scores, two sessions of the Speech-to-Text APIs were operating in parallel on the robot, one for each language used here (English and Italian). Through the APIs, the robot listened continuously for human speech and real-time transcripts of audio chunks from microphone streaming were yielded on demand. On a third thread, continuous and automatic query requests would detect spoken language automatically from the audio contents prompted by user speech. The detected language code (en-GB or it-IT) determined the transcript of which API (English or Italian) would be sent to the learning framework through the YARP communication input port. Consequently, the transcript input phrase received by the learning framework was closer to the ground truth of the spoken audio. The latency of this operation was insignificant; audio transcripts were sent and received in real-time.

Figure 11 Bridging the cognitive architecture with the Q.bo One robot via YARP communication protocol and Google Cloud Platform. The flow of information to and from the cognitive architecture and the robot is shown. On the left side, the robot manages speech recognition, automatic language detection and speech synthesis, based on the predicted language code from the user’s speech. It feeds information to YARP over SSH, which runs altogether with the cognitive architecture on a remote computer (right), responsible for the information processing at a neural level.

The verbal output was managed by a YARP connection output port and sent to the Text-to-Speech (TTS) API operating on the robot end. The TTS API of the correct language was called each time without fail, based on the predicted language code from user speech (en-GB or it-IT) that was generated in the opposite direction. This mitigated the slight limitation of the Google text-to-
speech API to synthesise speech in only a single language at a time. The robot would speak the response received from the learning framework aloud and display the same content on the screen for enhanced readability.

### 3.5.3 The Behavioural Experiments

Due to the bilingual corpus being relatively large, the live experiments were performed using a sample of 26 distinct types of interrogative test phrases for each language (52 in total). The phrases were randomly extracted using the criterion that none was repeated in both languages and such involving an eclectic range from the bilingual set. In turn, this enabled to assess the efficacy of the speech recognition and synthesis and the flow of information in the final robotic architecture.

The human user would probe the robot with random questions from the test and casually switch languages. The robot would initially attempt to detect the human’s face, indicated by a green blimp on its nose that turned blue when the robot was ready to listen. Throughout the conversation, it would try to maintain eye contact using head rotation towards the detected face. The robot could detect the language automatically and respond in turn in the same language in the appropriate context. The proposed technique of running simultaneous speech APIs on the robot to obtain the closest ground-truth transcript yielded satisfactory recognition scores. Only in three cases, the human had to repeat the question.

![Illustrative samples from the live experiments with the Q.bo One robot, handling bilingual dialogue with the human in English and Italian simultaneously.](image)

The robot answered correctly 24 out of 26 questions in English and 23 out of the 26 selected questions in Italian (Video 1), yielding an average joint accuracy of 90.4%. Given that the
framework has learned the questions simultaneously in their own ambient, it can preserve context. Let us take for instance the question “How old is your sister” (She is 7 years old). Should the system rely on translation, the literal phrase in Italian would be “Lei è 7 anni”, which is not valid (instead “Lei ha 7 anni”). Though intended for developmental complexity using a simple lexicon, the embryonic results demonstrate a considerable potential towards achieving the long-term goal of multilingual robots that convey germane contexts. Figure 12 illustrates samples from the bilingual dialogue with the robot during the live experiments.

3.6 Conclusion

The hypothesis that governed the research in this chapter was: can we devise multilingual cognition in a biologically sensible AI model congruent with how the bilingual brain is activated in response to language and how it resolves language conflict? If an artificial cognitive architecture simulates low-level brain processes and exerts biologically sensible learning principles and mechanisms for processing verbal stimuli (i.e., language), can it develop humanlike abilities of language production in ANY human language? How close can the AI model approximate to psychological evidence on the way humans naturally develop language? To argue, this chapter investigated the potential of a developmental and biologically motivated cognitive architecture to acquire and liaise simultaneously in multiple (4) human languages with intricate structure and complexity. The intent was to verify if the selected learning framework could learn languages jointly at the same degree of skill as learning each apart and handle potential conflict amid languages. The artifice was inspired by the literature on the natural organisation of information in the brain of early bilinguals who store information in the same area of the brain without specific distinction of the language (Abutalebi, 2001).

The extensive evaluation in a large corpus demonstrated robust generalisation competencies of the learning framework in the acquisition and disambiguation of poly languages even for coherent content. Notwithstanding, it also suggests that there might be a slight interference in the joint processing of complex and distinct semantics, reflected in the performance decay in the polylingual acquisition, which can be overwhelmed by augmenting the input samples and/or the number of neurons and interconnections. The observed cases of resemblance with code-mixing that were consistent with language syntax suggest that the neural representations of multiple languages might be activated in parallel, resulting in conflict. This potential language conflict was successfully resolved by the learning framework at worst case over 78% for all languages, where the types and number of produced errors were analogous to monolingual acquisition. As such, the study introduces open questions that require careful investigation to determine if performance depends on the size of learnable connections for a predefined architecture (but not on the sequential order in which languages are learned) and what is the confusion that can occur in poly language
acquisition from a brain-inspired angle. In-depth exploration of varying the learning samples and the focus of the attention on word relevancy required by the working memory in decision making could be useful to boost the framework’s performance. On a final note, endowing such cognitive frameworks in social robot companions can forerun polyglot robots that self-understand and communicate contexts in multiple tongues, without loss of meaning in translation, which can greatly affect human cognitive and interaction capabilities. In analogy with how bilingual human infants use their limited vocabularies resourcefully (Lanza, 1997), a gripping matter is to devise this skill in robots, which have restricted resources, either in the form of pre-trained knowledge and behaviour, or their current frailty to adapt to real-world constraints. This human competence is devoted to our conceptual development, which, as this thesis thrusts, is inseparable from language. To arrive at robots that use accessible resources dexterously should they lack direct experience or cannot quickly retrieve relevant knowledge from one resource, it is imperative to first shape their conceptual development. To anticipate, this assumption is suitably modelled and demonstrated through the research described in Chapter 5.
Chapter 4

From Human Language to Robot Behaviour

4.1 Context

This chapter introduces a novel methodology to address the symbolic mapping of language with the robot’s physical locale, to acquire the meaning of concrete and abstract linguistic notions. The artifice seeks to obtain a robot that is not reliant on a pre-trained knowledge about the task and the environment for anticipated scenarios but acquires such knowledge when brought to interactions with a human in an assumed real-world workspace and uses that knowledge directly to achieve a new goal or outcome (i.e., runtime learning). The methodology devised in the learning framework is extensively validated in a series of learning experiments with a humanoid robot, divided into perception and manipulation tasks. The tasks evolve at runtime with the help of human guidance using only language instructions. The robot explores the new workspace and learns the objects involved in the task along with a set of primitive actions enacted upon those objects that are dictated by the human, by self-mapping the visual and motor representations (i.e., meaning) to the human’s spoken utterances (i.e., linguistic terms). The human continues to intricate the task, adding more (abstract) instructions that require to combine several earlier-learned primitive actions, until the goal is reached. Human involvement is important for this cognitively plausible setting. The human guides and loosely supervises the online learning of the robot, which builds the complex task from the ground up in smaller steps only by following the human’s verbal instructions on the workspace. The methodology focuses on learning a task in developmental stages akin to child-training (and attainable from a non-programmer human tutor), rather than representing high-level actions as the desired goal state.

The research described in this chapter is published in Giorgi et al., (2021).

4.2 Prologue

4.2.1 Workaround the Visuo-Spatial Sketchpad

Passage 2.3.4.1 offered a comprehensive description of the analogy of the (ANNABELL) learning framework leveraged in this thesis with the theoretical working memory models. A central aspect that the learning framework has not conceived in its original form is the manipulation of sensory (non-verbal) stimuli, commonly the perception and the motor data in robotic contexts. The working memory insights suggest that our brain uses separate units to process such information of different domains (verbal, non-verbal), given theoretical and empirical evidence that the mind can perform a visual and a verbal task, but not two visual or two verbal tasks at the same time (Baddeley, 2010). Hence, the WM model that has originated from these claims makes a clear distinction between the phonological loop and the visuospatial sketchpad Baddeley & Hitch (1974). ANNABELL does not include a separate structure for this purpose, hence, refinements must be made to accommodate multimodal data in the framework. One way to tackle this would be to modify the architecture to introduce the visuospatial structure. Instead, the solution proposed here makes no amends to the
global organisation of the model but rather to the representation and binding of information from the multiple domains inside the framework (episodic buffer). The proposed method takes direct inspiration from the embedded–process model of the working memory (WM) in the Cowan (1995) perspective. Cowan recognises neither a phonological loop nor a visuospatial sketchpad (and nor an episodic buffer) but implicitly suggests that the functionalities of these structures are maintained by the focus of attention of his model and the executive control process that deal with the short-long-term stored information. In this view, the working memory is only a subset of an activated long-term memory, which, in turn, is an intermediate subset (short-term storage) of the long-term memory. Thus, without specific distinction of where does the multidomain information emerge from, the working memory holds such domain-specific information (phonological, visual) in the form of activated units. The focus of attention (called episodic buffer by Baddeley) bounds these activated elements in integrated units of information with chronological sequence. This claim has been supported later in time by the alternative perspective of Baddeley (Baddeley, 2010).

The proposal offered in this chapter is motivated by the postulate of Cowan and seeks to devise an artifice that binds domain-specific information from the linguistic and non-linguistic stimuli in the existing learning framework, without introducing structural changes. The proposed method (4.3.2) uses symbolic conventions for the internal representation of the non-verbal data and exhaustively trains the short-long-term memory units of the framework to bind and retrieve the information of each domain (phonological, visuospatial) readily and aptly.

4.2.2 The Competence for Runtime Learning

Before embarking on the artifice offered in this chapter, it is imperative to validate if the learning framework can accommodate online learning at runtime, i.e., from interactions occurring after its pre-training. The question that thrusts this investigation is: can it (the framework/a robot) acquire information on the fly on interactions with the environment that are not explicitly anticipated at design time and use this newly acquired knowledge directly to support those interactions?

The venture starts with the robotic system described in section 3.5.2, modelling a rather atypical learning: instead of leveraging on a background knowledge stored in long-term memory to resolve the human’s queries, gradually gain this knowledge from the interaction itself, during which the dictation of knowledge and querying on it happen side by side. Q.bo One robot is trained on only one occasion to name the category member of some animal. No other animal is known, i.e., memorised in long-term memory. At the end of this initial pre-training, the robot is brought into interaction with the human, who verbally dictates further information of close context at runtime, akin to a toddler learning from speech descriptions. Can the model successfully retrieve this lexicon when queried on it to pursue real-world interactions? Another appealing matter is what would be the robot’s counter to a question it has no explicit knowledge on. What follows leads to the ingenious intuition: if the robot is dictated the answer, will it amend its past decision and produce a new answer that successfully responds to the question?

The two questions pitched above are addressed in two scenarios with the learning framework as per the following method (Figure 13).

Scenario 1: Learning begins with limited explicit knowledge, involving only the affirmation “the python is a reptile”, permanently stored in long-term memory. The framework is appropriately trained to elaborate the interrogative phrase “what is the python” to build the valid output “a reptile”. In the runtime test, the human dictates more affirmations on other animals: the <animal> is a <class member>, where <animal> is an animal name (e.g., cat) and <class member> its
category (e.g., mammal). The human can control preferentially the size of the vocabulary that the robot learns, using natural language only. Next, the human queries the robot on the newly acquired knowledge and validates its competence to learn novel data from real-world verbal interactions.

Scenario 2: Learning precedes the runtime test of scenario 1. After training on known information (python), the learning framework is taught to manoeuvre the absence of it (favourable feedback in partially incomplete scenes). This is triggered by interrogating the robot with the sample phrase “what is the fly”, before learning the fly (i.e., the fly is an insect). The robot was trained to reply with “I do not know the <animal>”, where <animal> is the queried animal, which at neural level processing means futile information retrieval from long-term memory. A central competence to flag here is that the robot can self-handle the presence or absence of information: i.e., to the question “what is the python”, it replies “a reptile”; to the question of similar structure and context, “what is the fly”, it replies “I do not know the fly”. Next, the robot is verbally dictated the answer and trained to build a valid output. When re-queried on the fly, the robot overrides its past decision and replies “an insect”. Hence, scenario 2 validates the robot’s competence to adapt its behaviour as per novel data acquired from real-world verbal interactions.

In the test, the human initially dictated the categories for 20 novel animal names (e.g., the cat is a mammal). A further 10 animal names were used to query before dictating them, to which the robot successfully replied, “I do not know the <animal>”. The human dictated the answer verbally and queried the robot randomly on all (30) animals. The robot demonstrated robust generalisation properties, validating three salient competencies:

1. To learn novel data from real-world interactions occurring at runtime, and directly use this verbal experience to support the interactions.
2. To handle information retrieval autonomously and provide language feedback to the user for incomplete contexts.
3. To adapt to the changes in information in the real-world environment and self-regulate its behaviour to accommodate/exploit that information.

For examples of the live behavioural experiments, see Video 2.

4.2.3 ON-World Embodiment

Duffy & Joue, (2000) distinguish between a robot being only a controller with actuators and preceptors that performs in the environment without being part of it per se and the robot being part of that environment so that the direct interaction with the environment influences the robot's real-time learning, adaptation, and development in it. They define the former as ON-World or weak embodiment and the latter as IN-World or strong embodiment. These notions are also supported in the work of Sharkey and Zeimke, (2000). In the method described in this chapter, the robot is placed in the physical world and functions autonomously by means of appropriate elaboration from the coupled learning framework of the sensory inputs that situate the body of the robot in its internal map. In this sense, the robot is ON its environment. However, the next sections will articulate how starting from an initial set of motor primitives, the robot will autonomously generate new maps and representations of the world to actively interact with and influence the environment beyond its static internal representation.

A humanoid robot was selected following the intuitive reasoning that a physical body is necessary to act in the environment and that natural language as a communication interface is arguably a human attribute. The robot is an academic edition of the NAO robot, version 6 (SoftBank Robotics, 2021). NAO's motor skills are limited, however, its relatively low cost and computational effort make it suitable for the type of experiments performed here, which place a greater emphasis on cognition over actuation. The robot is coupled with two pre-trained state-of-art vision recognition systems, AlexNet (Krizhevsky et al., 2012) for object classification and YOLO (Redmon et al., 2016) for object tracking, using Tensorflow 2.0 and Keras. AlexNet was preferred over its latest higher-accuracy competitors, because of its faster training time and lower computational power requirement. It renders an acceptable classification performance for the object classes used in the perception learning experiments described here. Instead, the YOLO model was a better fit for the manipulation learning experiments to grant fast real-time tracking of objects and improve sensory inaccuracies. The aim was to maintain the external components lightweight and allocate the available computing resources to the training and validation of the large-scale learning framework. ANNABELL, the NAO robot and the smart vision are coupled through a client-server protocol. On the server side, ANNABELL runs on a remote computer and communicates in two directions with the robot via the robot's local network (Figure 14).
4.3 The Learning Methodology

The learning methodology draws on fundamental research aims.

1. How language can be appropriately mapped to the robot’s self-representation of the tactile environment to acquire the meaning (physical manifestation) of primitive (concrete) and higher-order (abstract) semantics.

2. How to model a cognitive robot whose knowledge is not conditioned by pre-trained explicit corpora for anticipated scenes, but rather continues to acquire (amend and expand) such knowledge when brought to interaction in unfamiliar workspaces and apply the knowledge directly to attain a novel goal or outcome.

In simpler words, the robot must learn conceptual meanings through the symbolic capacity to map language and environment (datum 1), scaling up perpetually at runtime on novel vocabulary that is not part of the initial learning corpus (datum 2). In this chapter, learning framework and robot are used interchangeably, and by system, it is meant the final end-to-end model including the robot, the coupled external components, and the learning framework.

4.3.1 Pre- and Post-Linguistic Primitives

The cognitive neural architecture is interfaced with the NAO robot, which receives spoken utterances from a human interlocutor through its internal speech recognition API and generates robot motor behaviours from visual and auditory stimuli (Figure 14).

The visual data are extracted from the real-time frames captured by NAO’s cameras and processed by the bridged detection/tracking external modules. These include directly observable instances (percept), such as objects, colours, shapes, etc. The motor data are primitive operations executed by the robot using elementary movements of the joints or body. They include only fundamental action primitives such as grasping, lifting, pushing, pulling, rotating, ..., which can be coordinated and combined in some way to generate more complex robot actions, at higher levels of abstraction such as to take/make/place/use/... something. Here, sensory data or sensory stimuli refer to the
percept and motor manifestation and comprise the non-linguistic domain. Transcripts of spoken utterances – auditory stimuli - comprise the verbal data (linguistic domain).

Percept and motor stimuli can be expressed in lexicalised form through a term. The linguistic terms are otherwise referred to as semantic form or form. The manifestation of percept and motor data in the physical world is the meaning of those forms. For example, the term dog (i.e., form) refers to the animal that is seen or the visual mental representation of the real animal (i.e., meaning). Similarly, the term grasp (i.e., form) describes the act of moving the joints to seize and hold an object (i.e., meaning). For both linguistic and non-linguistic data to co-exist and be manipulated jointly in the cognitive neural framework, there must exist some knowledge on the <meaning, form> pairs and form must be precisely discerned from meaning. Sensory stimuli must, therefore, be conventionally and appropriately represented, in such a way that they can be disambiguated from the verbal information stored and manipulated in the phonological loop. Hence, the phonological store now holds domain-specific activated elements bound in integrated units of information, as in the Cowan’s model of the working memory.

### 4.3.2 Binding Domain-Specific Information

To represent the sensory stimuli conventionally in the verbal-only working memory (learning framework), linguistic symbols are used. These are generated by prepending an underscore (_) to the respective term (i.e., form) (Figure 15). The newly created symbol uniquely identifies and pivots to the sensory stimuli received from the robot. This process of representing objects and robot motor experiences through human-readable linguistic symbols, here, is referred to as a symbolic representation of sensory stimuli. Predominantly, sensory data are often represented in the form of binary streams (or pixels) and joint angle vectors, however, given that the output is supervised, these representations are less human-readable during training and validation. Instead, lexicalised symbols are easier to interpret. Throughout this chapter, symbolic representation denotes that meaning (real-world sensory stimuli) is verbally represented as _form (linguistic symbol) and <meaning, form> pairs are manipulated as <_form form> pairs in the learning framework. For example, if the robot captures the image of a dog, the robot’s coupled vision system processes the image and generates the label _dog upon recognition, which is sent as input to the framework. Similarly, a trajectory movement of the robot’s joints, corresponding to an action primitive, say a grasping-like movement, is encoded as _grasp. These representations are uniquely paired to their semantic terms (form) and stored permanently in long-term memory (e.g., _dog dog, _grasp, grasp). Hence, the robot knows the link anchored between form and meaning, as well as how to discern them (Figure 15). When the robot hears the word dog or the instruction grasp it can retrieve the corresponding pair from the LTM (Long Term Memory) and, therefore, extract their respective meaning (_dog, _grasp), which drive a certain behaviour to the robot’s internal modules (vision/motor). The vice versa is also true: when the robot sees a dog or executes a grasp action, it can extract their respective lexicalised forms, through the memorised mapped pairs (Figure 15).

Encoding sensory information in verbal form (_word) as opposed to bitstream representations, any dataset labels (e.g., image net classes, n02119789) or joint angle vectors allows to:

1. Parse and elaborate visually presented data similarly to verbal stimuli and process both simultaneously, while properly disambiguating the information contained from sensory and that from language; therefore, this artifice is applicable in other complex cognitive architectures that exploit language, regardless of the larger embodied/grounding system.
2. Generate a \(<\text{form, meaning}>\) pair for every new instance (limited with fixed dataset labels) and retain a unique representation even when image bitstreams differ for the same semantic (e.g., different images of the same object); \textit{scalability and on-line open-ended learning}.

3. Appropriately represent knowledge and retrieval mechanisms, so that the multimodal information elaboration is a sole attribute of the learning framework and freestanding from the external coupled components or robotic platforms; \textit{optimisation}.

\textbf{Learning \(<\text{meaning, form}>\) pairs}

\textbf{Retrieving \(<\text{form}>\) from \(<\text{meaning}>\)}

\textbf{Retrieving \(<\text{meaning}>\) from \(<\text{form}>\)}

**Figure 15 Binding domain-specific information.** A) Sensory stimuli from the non-linguistic domain (visual, motor) are conventionally represented through a linguistic symbol. They are uniquely mapped to their respective linguistic forms (linguistic domain). \(<\text{form, form}>\) mappings are memorised permanently in the long-term memory (LTM). B) Meaning can be retrieved from form and vice versa.

The symbolic representations \(<\text{form}>\) can indicate a main classification (object, category, motor primitive), optional attributes – size, shape, colour (“big”, “round”, “red”), or spatial localisation in the scene (“left”, “right”, “centre”). Sequences of \(<\text{meaning, form}>\) (i.e., \(<\text{form, form}>\)) mappings set up a scene. These pairs are trained and memorised permanently in long-term memory (LTM). In the forthcoming sections (4.4.1.2, 4.4.2.2), it is discussed how they can be generated through a runtime human-robot communication to grant scalable and continuous on-line learning from situated interactions in the tangible workspace.

The high-level logic behind the proposed learning methodology is illustrated in **Figure 16** and can be divided into three parts, which are extensively explained in the upcoming sections:

i. Learning to map concrete words to the observable and tactile entities they refer to in the environment.

ii. Learning to represent abstract words that have weaker perceptual constraints in the environment, by extrapolating the representation of concrete words into (abstract) high-level complex behaviours.

iii. Generate continuous learning of novel concrete and abstract words adaptively, in and from task-oriented human-robot interaction at runtime.
Chapter 4: From Human Language to Robot Behaviour

Figure 16 The proposed learning methodology to integrate language with perception and action. The fundamental neural components of the learning framework are shown: the STM (Short-Term Memory), which handles multimodal input phrases and output production; the CE (Central Executive), responsible for the statistical decision-dependent processes on neural level processing (instead of a conventional IF-Else logic); the LTM (Long-Term Memory), which stores multimodal information that scales up during the runtime interaction. The CE can disambiguate between the task being learned or queried, along with the type of query requested in input. Should the instruction implicate concrete concepts, the robot self-maps the linguistic concept with the internal representation of the concept (visual or motor) and retains the association in long-term memory. Should the instruction involve abstract concepts, the robot uses linguistic experience and input from the user, to map the concept indirectly to earlier-learned (concrete) concepts that have a representation in memory. Learning of novel concepts (objects, actions) occurs at runtime, which involves learning feedback for cases of incomplete scenes. If a concept is queried before learning it, the robot hints the human without breaking the learning process. The human can teach the concept using verbal descriptions and live captured images or motor exploration, similar to child learning.

4.3.2.1 From Concrete to Abstract Concepts: Learning Snowballing

Learning concrete (observable) and abstract (lexicalised) concepts is modelled as follows:

1. Concrete/primitive concepts are directly represented as <form> pairs (concreteWord in Figure 16 used for convention), where form is the linguistic term of the real-world representation_form. These are stored permanently in long-term memory (_concreteWord_concreteWord) and adhere to the artifice of binding domain-specific information explained in section 4.3.2.

2. Abstract/higher-order concepts are indirectly associated with multimodal representations (language and actions), by extrapolating the representations of primitive concept guided by high-level human language explanations. Given the speculated continuum from concrete to abstract concepts, the latter can be constructed as a combined series of primitive (lower order) concrete concepts and are indirectly mapped to the physical scene (abstractWord_concreteWord1_concreteWord2 ...). This mapping is facilitated through human-guided language instructions and their visual or motor interpretation by the robot itself. Notice that abstractWord does not have a pseudo-sensory representation, i.e., _abstractWord, as it is not directly mapped in the workspace, but indirectly through other words/objects/actions.

Constructing the representation of higher-order (abstract) words from their lower-order composition creates a snowballing effect of the learning experience. Starting from action
primitives that are defined in the robot’s internal motor map, snowballing generates new internal representations & motor modalities, which enrich the robot’s behavioural praxis in the workspace. Therefore, the robot can adapt (to some extent) past its predetermined internal state.

![Diagram](image)

**Figure 17. The snowballing artifice.** Action primitives are programmed in the robot’s motor map as predetermined trajectories (grasp, lift, drop). The human guides the robot through simple natural language instructions to combine the motor primitives suitably to produce a new outcome. While elementary actions are directly hooked to a specific motor exploration, complex actions expressed by more abstract verbs (take, place, make, add) entail convoluted motor modalities and their meaning is indirectly mapped to a certain combination of lower-order actions.

An important merit of the snowballing approach proposed here is that generating new higher-order instances requires little to no programming but relies solely on natural language interactions. This is illustrated in Figure 17. The action primitives grasp, lift and drop are pre-programmed on the robot end in the form of predetermined trajectories. They are used as motor stimuli for the cognitive architecture, using the convention "<_form>", i.e., "<_grasp", "<_lift", "<_drop". The learning framework is suitably trained by means of mental actions to convert human verbal instructions to output symbolic representation sequences that trigger the robot’s sensory response. The human uses guided natural language (NL) explanations to teach the robot the meaning of a higher-order word.

*to take the cup*

  step 1 grasp the cup
  step 2 lift the cup

Action primitives (grasp, lift) have unique "<_form>" pairs stored in the LTM ("<_grasp grasp>", "<_lift lift>"). Higher-order actions do not (e.g., take). Moreover, the robot does not have a planned motion path for higher-order actions. The robot cannot convert take to _take, because an internal motor representation of the action “take” does not exist (in the robot’s motor system) and neither does a "<_take take>" pair in the long-term memory. With suitably trained mental actions, the robot exploits the NL explanation “to take the cup” and the step-by-step instructions, to converge the meaning of the verb take as the combined outcome of grasp and lift (the cup). Notice that the
object *cup* is fundamental to identify which primitives are involved with the verb *take*, in this context. If the referent changes, so do the interactions. In another context, the outcome of the verb *take* could be to *walk* and *board*.

**to take the bus**
step 1 walk to the bus
step 2 board on the bus

In Figure 17, the verb *place* would require threefold steps. Hereby, to place (a cup) involves *grasping*, *lifting*, and *dropping* the cup (on a table), or first *taking a cup* (1st high order) and then *dropping* it (2nd high order). Verbs with increasing order of abstraction would require further primitive steps in the sequenced toolchain.

### 4.3.2.2 On-Line Open-Ended Runtime Learning

After the initial pre-training, the human can choose to query or continue teaching new concepts. The flexible learning mechanisms and knowledge representation in long-term memory allow the robot to learn an unrestricted number of novel concepts and apply the newly acquired data directly, without re-training the whole framework. Each novel instance that is introduced after the training stage, is forthwith memorised by the robot, scaling up its explicit knowledge and self-regulating decision-making in future endeavours. The human-robot communication is conducted at runtime and learning is guided by natural language.

### 4.4 Robot Learning Experiments

The learning methodology was applied and tested in a series of verification experiments with the final robotic model in two task-oriented contexts:

1. **Perception tasks**, in which the robot learns objects in its ambience, maps the lexical terms for objects to their corresponding percept (physical representation) and engages in simple gameplay interactions with the human on these objects.
2. **Manipulation tasks**, in which the robot learns to use assorted objects in its workspace, maps concrete action words to motor actuators (tactile representation) and extrapolates further complex manipulation behaviours in the scene from the preceding low-level modalities.

#### 4.4.1 Perception Tasks

The human and the robot engage in simple interactions, which require that the robot solves queries on a set of (observable) animals. The only action executed by the robot here is pointing. There are three developmental tasks:

1. The robot is asked to recognise an animal by its name and category member.
2. The robot is asked to discern which between two animals in the scene belongs to a queried category member.
3. The human presents three animals and hides one without the robot looking; the robot must identify which animal is missing in the scene.

#### 4.4.1.1 Target Training Data

Through preliminary training using the binding domain-specific information method (**4.3.2**), the learning framework has permanently memorised in long-term memory (LTM) the representations...
of (only) three animals: dog, frog, snake. Part of the explicit knowledge are simple affirmations such as the dog is a mammal, the frog is an amphibian, and the snake is a reptile. This is the corpus used to train tasks 1-3. The training set is limited: the goal of the method is to grant adaptive behaviour and flexible learning of novel entities (test set) at runtime, i.e., when the robot is brought to interaction with humans in real scenes. In simpler words, the methodology can be understood as learning how to do something rather than learning to do something. This allows the robotic model to apply the same line of reasoning in unfamiliar endeavours without a priori learning of the corpus or the workspace.

In task 1, the robot is trained to respond to questions of type what is this and what kind of animal is this. An example is given below:

**Human:** what is this _dog (*shows an image of a dog*)  
**Robot:** a dog  
**Human:** what kind of animal is this _dog (*shows an image of a dog*)  
**Robot:** a mammal

The human’s utterance (what is this) is processed by the speech module. The robot observes the animal (e.g., dog) through the cameras and the coupled vision module generates a label, which is encoded with a symbolic representation (e.g., _dog). The combined output of both components is elaborated by the cognitive architecture at a mental level (e.g., what is this _dog). The runtime acquisition and information flow in the robotic model are thoroughly explained in section 4.4.1.2.

A central aspect of the runtime adaptive learning that is particularly useful in a real-world human-robot interaction is interactive feedback. Suppose that the human decides to query the robot on an animal it has not learned, e.g., an elephant. Given that the robotic model has no explicit knowledge of the animal, the memory retrieval mechanisms will fail (as explained in 4.2.2). To consider these occurrences, the robotic model is trained to announce that it does not know the answer. If humans are told the answer to a previously unknown query, they can successfully learn it for future use. The robot is then trained to manifest a similar behaviour. The human dictates the correct answer to an unresolved query, which the robot memorises for future endeavours. When re-queried, it will self-regulate its past decision I don’t know this to a correct answer, e.g., it is a mammal. This type of learning using language feedback has also been studied by **Twiefel et al. (2016)**. The authors have proposed a rule-based inference module, which can identify cases of inconsistency in the incoming sensory information from the current workspace configuration, reject invalid inputs and maintain a feedback loop to the user to change the uttered instruction.

The robot is trained to solve task 2, using questions of type which one is the <animal> and which of the animals is a <category member> (with same explicit lexicon). An example is given below.

**Human:** which one is the dog _dog _frog (*shows an image of a dog and of a frog*)  
**Robot:** the dog (*the robot points at the dog*)  
**Human:** which of these is a mammal _dog _frog (*shows an image of a dog and of a frog*)  
**Robot:** the dog (*the robot points at the dog*)

In task 3, the robot is trained to identify the missing animal by sight, as illustrated below.

**Human:** there are a dog, a frog and a snake  
**Robot:** you see _dog _frog (*shows an image of a dog and of a frog*)  
**Human:** which animal is missing  
**Robot:** the snake
It can be seen how, for this context only, the robot “interprets” the inferred concept of a *missing* object as something that is not present in its perceived workspace. At the end of this learning stage, the robot is brought to runtime interaction with the human in the workspace.

### 4.4.1.2 Runtime Object Learning

The purpose of perception tasks is not animal classification, which can be done satisfactorily with any object recognition or neural network classifier, but the interpretation of language instructions from visual stimuli. The study aims to demonstrate that the robot can acquire new words at runtime, self-map them in the tangible workspace and use this knowledge directly to generalise on unfamiliar endeavours. The robotic model must act autonomously without re-training on the new environment-driven or interaction-driven data and as such granting a degree of scalability and adaptability in novel workspaces.

This section describes how the robotic model can learn a new object by combining sight and auditory stimuli during the runtime interaction with the human. A learning example of the object *ball* is illustrated in Figure 18. It is modelled as a simple natural communication between the human and the robot alike child learning. The human shows the object to the robot (*percept*) and verbally dictates what the object is (*term*). The frames captured by NAO’s cameras are processed by the interfaced vision module, which performs a classification task, and the label is encoded in its symbolic representation (*_ball*). The human utterance (*This is a ball*) is processed by the internal speech module and transcribed to written content. The pseudo-sensory stimuli and the auditory stimuli are appended and the combined string (*This is a ball _ball*) is fed as input to the learning framework, which self-generates the paired representation <*_ball ball*> and stores it permanently in long-term memory (Figure 18). It is intuitive to ask why the framework extracts the content word *ball* and not the function words (this, is, a). The symbolic representation *_ball* is closest to the word *ball* among all words of the input phrase and as such it is retrieved to build the pair, resulting in only strongly perceivable objects being mapped to a lexical term.

![Figure 18](image.png)

**Figure 18** Object Learning. The robot is shown an object (ball) and verbally dictated its linguistic term (this is a ball). The classification label from the vision module and the recognised speech are jointly employed in the learning framework, which self-generates the association <*_ball ball*> that is permanently retained in long-term memory (LTM). The newly acquired mapping can be uniquely and directly retrieved from memory in future endeavours.

Each time an object of the same category [ball] is shown to the robot (without distinction between different types of), the sensory stimuli extracted from the robot’s vision will retrieve the internal memorised representation <*_ball ball*> (in the learning framework) and readily use it in task-accomplishment goals. Any other object not yet learned will generate a pseudo-sensory (symbolic) representation, say _object1, which does not have a mapped lexical term in long-term memory (the robot has not been dictated its label). Henceforth, memory retrieval will fail, resulting in the robot
successfully discerning between the learned and unlearned data. In the latter case, it self-initiates the learning feedback (the example of the elephant in section 4.4.1.1).

### 4.4.1.3 Target Test Data

The target test data are acquired gradually at runtime using the object learning method explained in section 4.4.1.2. The test aims to verify (a) if the robot can adaptively scale up its long-term knowledge at runtime without altering what was known previously, (b) its certainty to infer if the current knowledge is consistent to resolve a query and (c) the competence to self-regulate its behaviour autonomously in the workspace, when the information about the workspace changes.

The robot is taught the new objects directly from verbal utterances and live captured images using the same setup demonstrated in Figure 18. The test data are collected as follows. The NAO robot and the bridged external modules operate on the client side and communicate bi-directionally with the learning framework on the server end (Figure 14). The human shows NAO a new animal from a printed image. The camera-captured frames in the robot’s visibility area are classified directly and encoded to pseudo-sensory symbolic representations. The human verbally dictates the animal. The outputs of the coupled external devices are loosely monitored by the human teacher, to pre-filter sensory-based errors in the learning samples, so that pairs are generated correctly (see 4.5). The learning framework elaborates the combined input (sensory representation – semantic label) to generate the association, resulting in the quick population of the long-term memory with novel <animal animal> representations. When sufficient animals are learned, the human initiates an execution phase (queries 1-3) on the newly acquired data (4.4.1.1). The results are disclosed in section 4.6.

### 4.4.2 Manipulation Tasks

The robot is tasked with manipulating objects and executing more complex behaviour from natural language instructions. It learns to map instructions expressed by concrete words to its own body and motor exploration and next self-generate convoluted motor modalities from the initial motor map of action primitives to counter more challenging requests. The robot conduct is built gradually from the ground up, in which the human first guides the robot to map concrete verbs to their direct tangible meaning by performing an elementary motor exploration (i.e., primitive action) and then introduces verbs of increasing abstractness that are not pre-programmed in the robot's motor map, whilst suggesting what step-by-step manipulations are required to achieve the meaning of those verbs. In future contingencies, the robot can autonomously dismantle similar complex tasks into simpler composite steps (continuum) following the same line of reasoning. This artifice of using active language instructions to guide the self-extrapolation of low-level action primitives into more complex tasks allows the robot to converge the meaning of abstract verbs such as make, use, take,... in its workspace.

#### 4.4.2.1 Target Training Data

In the training stage, the robot is taught to prepare tea using three objects: a mug, water, and a teabag. From an initial pre-training, the robot has learned the set of requisite tools for the task and three elementary actions, grasp, lift, drop. Section 4.4.2.2 describes how these language-to-action mappings are acquired at runtime. The tools are placed on a table-top setting: therefore, the table is part of the workspace.

Training the task is conducted as follows. It is assumed that the robot has some basic motor skills in its internal representation: NAO in record modality (enabled in a real robot) is trained via direct
motor exploration of the body and joints the somatic ability to grasp, lift, push, pull, ... (action primitives). The joint movements are regulated to produce a trajectory formation that resembles an action. The trajectories are encoded in symbolic representation, in which each linguistic symbol represents the sequence of joint angle vectors along the predefined trajectories (e.g., _grasp is mapped to a respective grasping-like trajectory in the robot’s internal motor map). This convention enables proper training of the long-/short-term components of the learning framework in an on-line scale-up vocabulary. Although the series of motor primitives are pre-determined and static, the learning applied here enables the robot to generate and learn autonomously the mappings between the action and the internal representation corresponding to the action (the language-to-action link is not manually forced). Moreover, the initial static motor representation of the robot is gradually refined with novel representations (attribute of language), when the robot extrapolates the motor map to build more complex behaviour that is not predefined in its body.

Training implicates also learning to manipulate the first object mug using the snowballing learning approach explained in section 4.3.2.1. It is assumed that the robot has already pre-learned the mug (4.4.1.2) and what it means to grasp. When the human utters the instruction grasp the mug, the learning framework retrieves the symbolic representations of _mug and _grasp, while the robot self-retrieves its respective trajectory from its internal map and executes the action upon the object mug. At the end of this learning stage, the robot is brought to runtime interaction with the human in a joint workspace. Section 4.4.2.2 details how the language-to-action link is self-generated from the runtime interaction.

### 4.4.2.2 Runtime Action Learning

This section describes how the robotic model learns language-to-action mappings at runtime. These mappings must be accurately represented in long-term memory to be retrieved in task-solving and to establish a successful link in the robot's internal motor representation.

The long-term memory is populated with novel language-to-action pairs, involving verbs that describe primitive actions (language) and their direct tangible meaning extracted from the robot’s motor stimuli (action) (Figure 19). The runtime action acquisition assumes that the robot must have some innate basic motor skills, to begin with. The human triggers an exploration on the robot (e.g., grasping trajectory) and dictates its meaning (e.g., this means to grasp). The internal representation of the action from the robot’s map (motor stimulus) generates the encoded symbolic representation (_grasp) and the combined input (this means to grasp _grasp) is fed to the learning framework to self-generate the <form, meaning> (e.g., _grasp grasp) pair, which is permanently memorised in long-term memory. (Notice that the linguistic label grasp is not part of any pre-determined corpus and its mapping to the motor modality is not pre-programmed but learned from experience and on-line interaction with the human and the ambience). The aim is to explore how these mappings are then retrieved and used in more complex explorations that do not have a direct association between an instruction and its manifestation but involve several self-generated motor modalities (higher-order instances in the snowballing learning).

Similarly, Hinaut et al., (2014) have explored the acquisition and production of grammatical constructions. In their method devoted to action performing tasks, the meaning of a random action is obtained by probing the robot to generate the action using some objects in the workspace. The human observer utters a meaningful sentence that can be used to command the robot to perform that action. Thereby, <sentence, meaning> pairs are created to populate the database and are used during human–robot interaction to instruct the robot to perform actions. One major difference with the method described in this chapter is that here, this type of mapping is used for primitive actions.
only, whereas full sentences of meaningful verbal instructions are generated autonomously by the robot with loose guidance from the human, by extrapolating its primitive self-representations (snowballing Figure 17). Otherwise speaking, the robot learns to converge sentence meanings gradually rather than a direct sentence-to-action mapping. In this manner, abstract concepts are not associated with a direct predetermined internal representation but are moored to their constituents of stronger perceptual constraints with motor experience and tangible referents (continuum).

![Figure 19 Action Learning. The human triggers a motor exploration on the robot (e.g., grasp) and dictates its meaning (e.g., this means to grasp). The robot retrieves the action trajectory and generates the respective pseudo-sensory representation, which is combined with the recognised speech and processed jointly in the learning framework. The WM of the framework self-generates the association "_grasp grasp", which is permanently retained in long-term memory to be readily accessible for future use.]

### 4.4.2.3 Target Test Data

The generalisation competencies are probed by placing the robot in a novel workspace and/or introducing instructions that involve the use of different tools to obtain a new outcome. There is no limitation in the number and type of novel objects that can be learned, however, the toolchain test data acquisition has two main constraints: (i) a set of primitive action trajectories must be initially predetermined in the robot so that verbally expressed actions have an internal representation in the robot’s motor map (assumed somatic skills); (ii) the types of phrases used to request (any) instructions must be similar to those trained, follow the same three-step sequence (1-3), and the final task must use exactly three objects with three manipulation steps for each object.

The test data are collected as follows. The robot is taught the requisite tools for the task, using runtime learning of real objects as explained in section 4.4.1.2. Next, the long-term memory is populated with new language-to-action mappings of simple motor primitives using runtime action learning explained in 4.4.2.2. The human starts an active step-by-step learning snowballing at runtime, to evoke the symbolic capacity to associate human language instructions to their meaning, involving new manipulation verbs with increasing abstractness in each level. The robot gradually learns to compose higher-order (abstract) concepts from lower-level actions (forward direction). The phrases used in this runtime learning should follow the structure illustrated below:

*To 1\textsuperscript{st} higher-order manipulation verb 1* the *object*

- Step 1 *primitive action\_1* the *object*
- Step 2 *primitive action\_2* the *object*

*To 2\textsuperscript{nd} higher-order manipulation verb 2* the *object*

- Step 3 *primitive action\_3* the *object*
Figure 20 Manipulation verbs that relate to primitive (concrete) actions are directly mapped to their internal sensory representation and the associations \(<\_action\ action>\) are permanently maintained in long-term memory. Human instructions that involve such actions (“step X do something”) are autonomously elaborated by the robot, which converts them to low-level symbolic sequences (representations of concrete objects and actions). Human instructions that involve abstract actions that do not have a direct representation in the workspace and, hence, not an \(<\_action\ action>\) association in memory, are extrapolated by logical linking of their lower-level constituents (indicated in steps). Two low-level action primitives are implicated in a 1st order of a high-level sequence. Higher-order manipulation verbs are generated by adding and combining duly more steps, in an incremental toolchain (2nd high-level sequences, 3rd high-level sequences, ...).

The robot starts an exploitation phase for phrases of type step x \(<\_do\ something>\) and memorises the output permanently in long-term memory. The output (symbolic representation of the object, symbolic representation of the action) triggers the robot exploration in the workspace on the object. The symbolic representation of the object \(_\text{object}\) prompts the YOLO module to track whether the referent object is in the workspace. If the object is recognised, the symbolic representation of the action \(_\text{action}\) retrieves the trajectory of the internal representation of the action from the robot’s map. The robot executes each of the steps 1-3, before the human introduces more complex instructions of the following structure.

\(<\_1^{\text{st}}\ \text{higher-order manipulation verb 1}>\ \text{the } \langle\text{object}\rangle\>
\(<\_2^{\text{nd}}\ \text{higher-order manipulation verb 2}>\ \text{the } \langle\text{object}\rangle\>...\)

For instance, if the robot has been dictated how to procure the bottle in two steps, step 1 \textit{grab the bottle}, step 2 \textit{lift the bottle}, a higher-level instruction would be \textit{Nao take the bottle}. The learning framework decomposes (backwards direction) the higher-order concept \textit{take} into the respective primitive concepts. It is important to notice that, only for this scenario the outcome of the command \textit{take} will be to \textit{grab} and \textit{lift}, because the object of reference is a \textit{bottle}, which determines the step-by-step lower-lever actions that are extrapolated.

4.5 Handling Sensor-based Uncertainty

Figure 21 illustrates how the integrated robotic model handles sensory inaccuracies of speech and vision recognition and what is intended by the accuracy of the learning framework alone.

In the\textit{ perception tasks} runtime tests, the robot receives two sensory stimuli (Figure 21). The transcript from user speech and the classification label of AlexNet are saved in separate strings.
To obtain satisfactory input samples closer to the ground truth, potential sensory errors are loosely pre-filtered by implementing a small idle time (~3 seconds) at the end of the speech and vision systems. This is done solely to enhance hardware limitations but does not cause significant latencies in the real-time interaction with the robot. During the idle time, the human supervisor monitors the sensory outputs. Should there be a speech misrecognition or object detection error, the teacher repeats the instruction or shows the printed image again. The respective output strings are overwritten automatically. The compound input for the learning framework is generated by appending the contents of both strings (a transcript and a classification label for the detected frame). The framework handles the input samples autonomously. Although pre-filtering helps to obtain close-to-ground-truth samples, it cannot eliminate all the sensory-dependent errors, while maintaining real-time communication. Therefore, some of these samples will arrive at the input of the learning framework. The transcripts of the uttered phrases include content words (nouns, verbs, adverbs) and function words (this, the, a). The framework can overcome inaccuracies when the recognition error is in the function words, as they do not have a mapping to a physical referent that is exploited during decision-making. Instead, content words are central to learning as they carry the semantic meaning of their real-world counterparts. Should there be a misrecognition of spoken content words, the learning framework will fail to create or retrieve the language-to-percept mappings. In these cases, the human teacher ceases and repeats the experiment.

![Diagram](image)

**Figure 21** Handling sensory imperfections during the runtime collection of target test data. In Perception Tasks (A), the framework uses a combination of user speech and a vision stimulus recorded by NAO’s cameras and classified by AlexNet. To obtain ground-truth sample inputs at the learning framework, the human tutor lightly filters potential sensory uncertainties. The human re-captures the input in the event of a speech (content words) or vision misrecognition. The accuracy is only tested at the ends of the learning framework (over the end-to-end robotic system). Similarly, in Manipulation Tasks (B) the framework receives a pre-filtered compound input of speech samples and trajectory information from the robot’s motor map. The robot’s motor information is collected and self-mapped to its symbolic representation without human supervision. The motor stimulus is critical for the construction of the test toolchain from low-to-high level instructions.

Typical errors occurring during the pre-filtering and those due to imperfect input samples, are a result of the interfaced external components and are not calculated in the total accuracy of the
learning framework. With that said, the accuracy of the learning framework alone is defined as in equation (2):

$$\text{accuracy} = \frac{\text{correct output sequences produced by the model}}{\text{total requested number of error-free and error-mitigated input samples}}$$  \hspace{1cm} (2)

where error-free samples are those that do not include any sensory uncertainty (misrecognition of the speech and visual modules) and error-mitigated samples are those input phrases, which do not contain misrecognition on the content word classes but only on the functional words.

The test acquisition toolchain during manipulation tasks has a stronger constraint for high-accuracy transcripts (Figure 21). The human must repeat the instruction and/or the experiment should content words be misrecognised. However, this does not break the toolchain. Only the intermediate failed command is repeated without ceasing and restarting the entire test. This process is smoothed using step indicators to maintain sequence order (step 1, 2, …). While this poses a slight limitation from a realistic angle of natural communication, it renders the method more robust to errors. Omitting such indicators does not affect the performance under perfect conditions (perfect samples) but produces cases of erroneous outputs when intermediate failed commands are repeated, hence occasionally breaking the test chain. Abstract content words affect the accuracy to the same extent, as they help connect the meaning of their composite lower-level primitives. During decision-making, the learning framework identifies that there is a resemblance in the abstract word received in the input instruction with the abstract word used by the teacher to guide learning that is stored as a goal (goal stack). This ensures that only the intended abstract words taught by the human are decomposed in lower-order mappings (and not any abstract word).

Motor data are fundamental for the acquisition of new primitives and the test toolchain. Their extraction is performed autonomously and solely depends on the robot’s internal representation and mapping competence. Should the robot not retrieve the trajectory data correctly, the learning framework will fail to process the information about the action, given that it lacks the requisite pseudo-sensory representation from the motor system. The mapping of the action semantic and its internal motor representation is done without human supervision. The mappings are necessary for the composition (forward) and decomposition (backward) of actions.

### 4.6 Results

The datasets generated for the work explained in this chapter comprised a limited lexicon (23 phrases in total, Table 4, Figure 23). The test sets are acquired at runtime by populating the long-term memory of the robotic model with new object and action representations from the workspace. The robot continues to learn objects, actions, and language past its pre-trained stage from a natural interaction with the human.

#### 4.6.1 Perception Tasks

Starting from a limited training experience, the robotic model is habitually probed with novel test data acquired from the tangible workspace. In the test, the robot first learned the language-to-percept mapping for ten (10) new animals from printed images. The phrases used to query the robot on developmental tasks 1-3 (4.4.1) are given in Table 4. The accuracy demonstrated in the test set is illustrated in Figure 22, where accuracy is measured using the calculation in (2). With correct output phrases are intended the valid verbal answers produced by the robot and cases of appropriate behaviour without language production, e.g., pointing at the correct animal. The overall accuracy for all language-to-percept tasks (~750 phrases) was 94%. Notice that this
accuracy represents the performance of the learning framework alone, where the sample inputs are perfect or insignificantly imprecise (section 4.5).

Table 4 The corpus of phrases used to train and test the robotic system the capacity to map language to perception. The test lexicon was generated at runtime using printed images of animals and speech descriptions.

<table>
<thead>
<tr>
<th>Task 1</th>
<th>Sample training phrases</th>
<th>No. of training phrases</th>
<th>Sample test phrases</th>
<th>No. of test phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object learning</td>
<td>_dog dog</td>
<td>3</td>
<td>_bee bee</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>_frog frog</td>
<td></td>
<td>_elephant elephant</td>
<td></td>
</tr>
<tr>
<td></td>
<td>_snake snake</td>
<td></td>
<td>..</td>
<td></td>
</tr>
<tr>
<td>Task 1</td>
<td>What is this _dog?</td>
<td>1</td>
<td>What is this _bee?</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>What kind of animal is this _dog?</td>
<td>1</td>
<td>What kind of animal is this _bee?</td>
<td>10</td>
</tr>
<tr>
<td>Task 2</td>
<td>Which one is the dog _dog _frog?</td>
<td>4</td>
<td>Which one is the bee _bee _bear?</td>
<td>2 \cdot P(10,2) = 180</td>
</tr>
<tr>
<td></td>
<td>Which one is the frog _dog _frog?</td>
<td></td>
<td>Which is a mammal _bee _bear?</td>
<td>2 \cdot P(10,2) = 180</td>
</tr>
<tr>
<td></td>
<td>Which is a mammal _dog _frog?</td>
<td>4</td>
<td>Which is an insect _bee _bear?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Which is an amphibian _dog _frog?</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task 3</td>
<td>There is a dog, a frog and a snake You see only _dog _frog Which animal is missing?</td>
<td>6</td>
<td>There is a bee, a bug and a whale You see only _whale _bug Which animal is missing?</td>
<td>2 \cdot P(10,2) = 180</td>
</tr>
</tbody>
</table>

TOTAL 19 750

To test the language feedback implemented during learning, half of the instances acquired in the runtime test (i.e., 5 out of 10 animals), were queried before learning. In all cases, the robotic model would initially announce that it did not know the answer. The human would verbally dictate the answer, which was used by the robot to autonomously self-regulate its past decision and respond correctly when queried again in future endeavours. An example is given below:

**Human:** what is this _swan (1st time query)
**Robot:** I do not know this
**Human:** this is a swan
**Human:** what is this _swan (2nd time query)
**Robot:** a swan

Play Video 3

Examples from the live behaviour of the NAO robot learning new animal names and classes (runtime object learning) are demonstrated in Video 3.
4.6.2 Manipulation Tasks

To generalise in a close context to the task of preparing tea, two new goals were defined: make me a juice and make me a toast. The former involves the tools glass, knife and orange and the latter bread, cheese and ham. The objects and elementary actions to manipulate them are initially learned at runtime to construct the test set.

The teacher initiates a snowballing learning with the robot, by guiding it verbally to execute actions of different and increasing complexity in the workspace in incremental steps, e.g., to place the glass on the table in three basic steps (grasp the glass, slide the glass, release the glass), to use the knife on the orange (grab the knife, lift the knife, cut with knife). The robot does not need any retraining to autonomously execute the spoken commands. Instead, by means of its experience, it converts language to pseudo-sensory representations that are mapped by the robot to the real sensory data. The higher the abstractness of the action indicated by the verb, the more the steps that are extrapolated to build higher-level executions in the toolchain. Let us consider how the system learns to use a knife, after having procured the knife (grasp and lift). The human verbally instructs the robot that to use the knife the next step is “step 3 cut with knife.” The semantic word use will be self-mapped to the representation of the constituent primitives, grasp, lift, cut. When the user requests to use the knife on the orange, the robot will execute the three lower-order primitives one at a time (clean start) or continue to elaborate the preceding order in the toolchain with an additional step, if the instructions of 1st and 2nd order follow one another (Figure 23). In this manner, although learning begins from a static internal representation of the robot, its basic motor map is adaptively used to construct and generalise to novel actions of increased abstraction, which are neither pre-defined in the robot's motor representation nor anticipated at design time but introduced through language.

**Figure 22** The accuracy rate for perception tasks. The tested vocabulary is not anticipated at design time but acquired preferentially during runtime interactions with the human. The human probed the robot with similar phrases to those trained involving the animal names dictated during the interaction. Without re-training the framework on the scale-up corpus, the robot could generalise with an overall accuracy of 94%, demonstrating a robust competence for symbolic mapping, open-ended learning of novel instances not specified in the initial pre-training, and reasoning on assorted interactive tasks.
Figure 23 The tested vocabulary is acquired gradually at runtime. The robot learned novel objects, actions and mappings, which were retained in memory in a readily accessible state for task-solving. It could construct grounded sequences for high-level (abstract) actions by transferring low-level constituents across multiple level domains and integrating language with the physical body. Language facilitated the learning of abstract concepts, which were indirectly self-mapped to their meaning by linguistic experience and their relation to earlier-learned concepts of lower levels. Starting from a simple motor map, the robot refined its motor modalities by creating new action representations aided by language in close liaison with the human and the workspace and learning to solve complex tasks by doing impromptu. The robot would converge the meaning of (abstract) instructions elicited from the current scene (available tools) and the outcome of the interaction (make – a tea, a toast, or a juice).

For the task “make me a juice,” the outcome generated by the robot was:

**Human:** Nao make me a juice  
**Robot:**
_-glass _grasp _slide _table _release  
_-knife _grab _lift _orange _cut  
_-orange _grasp _pickup _glass _squeeze

For the task “make me a toast,” the outcome generated by the robot was:

**Human:** Nao make me a toast  
**Robot:**
_-bread _grasp _lift _table _drop  
_-cheese _grasp _lift _bread _drop  
_-ham _grasp _lift _cheese _drop

It can be seen from the tested scenarios that the meaning of make depends on the outcome of the requested instruction (e.g., juice or toast), where the outcome involves employing a set of tools (e.g., glass, knife, orange for juice, bread, ham, and cheese for toast) and recalling a series of step-by-step operations acted on these objects (e.g., grasp, lift, release, cut) (Figure 23). The conducted experiments demonstrated a robust competence of the robotic model both in the adaptive learning
of new instances at runtime and in the successful execution of the two novel tasks, *make me a juice* and *make me a toast*, from little learning and in close liaison with the human and the workspace.

![Image of NAO robot preparing tea](image)

**Figure 24** Demonstration of the NAO robot preparing tea with a mug, a bottle, and a teabag.

### 4.7 Conclusion

This chapter coveted to address the research hypothesis: *How can a cognitive model develop the symbolic capacity to associate internal mental representations of real-world phenomena (meaning) with their labels (language)? How do abstract words take their meaning in the tangible environment? Can a model learn novel praxes from the workspace constraints and in interaction with other organisms in it, which can be used directly for problem-solving without re-training the entire model de novo?*

To confer the above premises, this chapter discusses a novel proposal to map language to robot perception and behaviour, by implementing adequate procedural, knowledge representation and memory retrieval mechanisms to the brain-inspired learning framework for successful elaboration of multimodal information from sensory and linguistic domains (extending its original capabilities intended for verbal processing alone). Here, the series of motor primitives are initially pre-selected, and the learning framework is then allowed to self-learn the mapping between the action semantic and the internal representation corresponding to the action. In contrast, [Yamada et al., (2016)](Yamada et al., 2016) investigate a more complex representation learning of these mappings, in a topological structure
of word meanings and their compositionality applied to a simple task. However, the artifice offered in this chapter to represent the real-world sensory data conventionally through human-readable lexical symbols can be justified with the appropriate training of the long and short-term memory units of the architecture to acquire a scale-up vocabulary of concepts, at the cost of a guided representation of sensory stimuli. A notable aspect of the methodology is the adaptive learning at runtime from novel data collected in a rather natural way, with direct teaching from verbal explanations and live captured images of the situated workspace, alike child training. The robotic model could execute complex tasks having no initial knowledge on the workspace and the task but acquiring such knowledge when brought to interaction with the human out of its initial trained stage. The learning property modelled here to construct new practices from preceding experiences, linguistic or not, generates a snowballing behavioural repertoire for the robot, starting from low-complexity modalities that convolute preferentially in increasing levels of abstractness.

One limitation of the proposed method is that the sensory data from motor primitives were defined and collected rather artificially, by first defining the actions on the robot using fixed trajectories. The actions performed by the NAO robot were intended for demonstration only, placing greater emphasis on cognition over conduct. In the long sight, this suggests the need for significant improvement, for instance, coupling the learning framework with external structures that allow for action generalisation on the robot end. Moreover, instead of triggering an exploration on the robot to learn the meaning of manipulation verbs, the robot could learn directly from the human demonstrating the action at runtime. By keeping the cognition inside the framework, sophistication is facilitated given the flexible coupling of the framework with any external module that can upsurge the final robotic agent. On the pitfall, the artifice proposed currently does not allow to investigate how the robot could naturally learn to predict, amend, and produce trajectories of angle joints according to the scene contextualisation and self-build appropriate internal representations.

The performance and stability of the integral robotic model from start to end is an important consideration. The pre-filtering of sensory input used here granted rather high robustness even under noisy laboratory environments (poor lighting, background noise), albeit could introduce a limitation for real-time real-world applications with human users. An important merit of the training methodology is that the errors that are not eliminated by the pre-filter do not greatly affect the continuation of tasks during the toolchain acquisition. The robot could continue to build (forward direction) and decompose (backward direction) higher-order concepts appropriately to handle the requested instruction, without repeating the entire experiment, but only the failed commands in the toolchain. The performance of the integral robotic model can be boosted remarkably by coupling more sophisticated technical shortcuts or error management methods.

As a final note, the methodology proposed here mitigates one typical problem in intelligent learning systems. In a classical neural network, a novel class of data cannot be added without re-training the whole network. The neural network can learn only if all the elements of the classes are part of the training set. Instead, the learning framework was suitably trained to learn adaptively new classes or new elements of classes, without training the entire model de novo. The procedural mechanisms of elaborating information by means of mental actions grant the adaptive behaviour and support open-ended incremental learning, which is a desirable goal for robotics and a novel outcome of the research described in this chapter with respect to the original framework of Golosio et al., (2015).
Chapter 5

High-level Cognitive Modelling in Robots

5.1. Context

This chapter that marks the culminating aim of the thesis marshals cognitive modelling of several high-level phenomena in a series of concept-like behaviour experiments with a humanoid robot built-in the considered learning framework. The robot acquires humanlike cognitive capabilities like learning, categorisation, abstraction, and cognitive control demonstrated in three experiments. The first experiment sets the foundation for the next two: in this trial, the robot learns the concepts and their attributes at runtime and produces interactions that depend on its developed high-level cognitive competence. The second experiment explores the robot’s decision-making process in response to linguistic stimuli that triggers new inferences about the event and outcome of the interactions. The third experiment investigates the effect of stimuli from different human languages in high-level phenomena during problem-solving, how the robot can bind information from different linguistic sources to resolve incomplete scenes and how language can trigger the same mental states as the motor experiences, to transpose those experiences in a new linguistic environment where they are not directly trained, and concepts are not explicitly learned.

Here concepts are regarded from a cognitive psychology angle. They can be represented via purely perceptual attributes (e.g., sensory stimuli), expressed in words, or be majorly abstract (e.g., ideas, counterfactuals). The problem of concept understanding is viewed from the angle of learning sciences research (Sawyer, 2005), which consent that understanding is demonstrated if the learner can:

C1: Identify examples of the concept that are subject to a high variation of non-defining attributes (e.g., a black swan classifies as a swan, regardless of the colour variable).
C2: Distinguish exemplars (an example of the concept) from close non-exemplars (something that is not an example) by assessing their significant attributes (e.g., a dolphin has significant attributes of a fish, but it falls in the category of mammals due to its criterial attributes).
C3: Maintain these abilities in novel endeavours that were not presented when first learning the concept.

The research described in this chapter is affiliated with publication (i).

5.2. Prologue

Before detailing the proposal of this chapter, this passage articulates some fundamental notions on high-level cognitive phenomena that are needed to understand the type of behavioural experiments and results obtained here. These notions have been previously introduced in section 2.2.2.1.

Categorisation

To categorise means to produce a given behaviour A when receiving a certain class of stimuli and another behaviour B when perceiving another class of stimuli (Miorilli & Parisi, 2011, pp. 303, p.

3 C1-C3 refer to criteria 1-3 of understanding.
2). It is the required competence that enables an organism to distinguish the first class of stimuli from the second. To conclude whether an organism/agent is categorising objects, actions, or experiences, it must exhibit a differential behaviour for the different instances: if the organism produces similar behaviour for some experiences (objects or actions), it is highly probable that they belong to the same category and to different categories otherwise. Categorisation can be, without a doubt, purely perceptual (no language). Yet, psychological evidence postulates that categorisation in humans often results from the production of different words that describe instances. Mirolli & Parisi, (2011) illustrate this with the following example: when a person hears yellow car, yellow apples, yellow flowers, the human brain produces a pattern for these objects in response to their joint word “yellow” that represents these objects as similar (in some sense), and different from the other classes of red car, red apples, red flowers. This categorisation results from the produced words “yellow” (same category) and “red” (different category). This can be validated even in the absence of colour perception (or if colour were to be replaced with more abstract words that cannot be perceived directly through senses, for instance beautiful - car, apple, flowers, ugly - car, apple, flowers). Therefore, humans produce a considerable number of categories in response to language over only those produced in the absence of language.

**Abstraction**

Abstraction is the cognitive ability to extract the features that are different among some entities when classifying them as the same thing and the features that are similar among some other entities when classifying them as different things. For instance, when classifying red car and yellow car in the same category, one must abstract their assorted colour; when classifying them into distinct categories, one must abstract the fact that they are both cars (shape, type). Abstraction is central to categorisation. In the forerunning example of yellow car, yellow apples, yellow flowers and red car, red apples, red flowers, the abstraction that led to categorisation was promoted by language. When labels are attached to categories, the abstraction ability is strengthened, and more categories are therefore built in response to language. As abstraction is an attribute of language, language has the overarching ability to promote the formation of new (more abstract) categories recursively.

**Voluntary Control**

Linguistic stimulus renders some aspects of experience more salient than others or draws more attention to some aspects while ignoring others, which changes the way the organism perceives the world and the interactions occurring in it. As a result, the organism (or agent) can change its decisions or behaviour in the environment. Assume that an agent has categorised some objects in the same category and has learned to produce some behaviour for that category. If a new linguistic stimulus changes the categorisation decision of the agent (objects grouped to distinct categories), the produced behaviour will change autonomously in response to the new categories. Language thus mediates voluntary control.

These phenomena and their manifestation in the presence of language are appropriately modelled and demonstrated in the to-follow methodology sections.

**5.3. Robot Concept-like Behaviour**

The behavioural experiments are designed around the problem of concept attainment, which is the act of using the discriminable features of an instance to anticipate its significant identity (Bruner & Austin, 1986). The experiments seek to verify the three criteria of concept understanding (C1-C3) and to demonstrate the manifestation of the high-level cognitive phenomena described in 5.2. During decision-making, the robot will seek to place an instance in a proper category niche, by
reasonably screening its salient attributes to infer with certainty if the concept is germane in the context and determine the proper course of actions. Inferences on events are better made within a context, which imposes specific constraints on what an event is, allowing the robot to assess the concept appropriateness. The categorisation ability is then tested by consistency (Bruner & Austin, 1986) in three concept-like behaviour experiments. The learning experiments were conducted with the learning framework described in 2.3.4.1 coupled with a PR2 robot (Willow Garage, 2022) in a virtual environment (Cyberbotics, 2022). Central enhancements are made to enable the learning framework to learn from one occasion and demonstrate high-level cognitive phenomena. The robot was initially guided to learn, categorise, and abstract the concepts in a foundation task and is later transferred to two novel convoluted scenarios for critical validation of its conceptual development. The robot is probed with solving tasks via inferences on lexicalised attributes that are not directly observed but allow it to make non-trivial generalisations. In a nutshell, the scenarios aim to verify how linguistic properties affect the robot’s abstraction, reasoning, and voluntary control, and how earlier-learned experiences are recalled opportunistically to novel tasks under new defining constraints.

5.3.1. Foundation concept-learning task: learning by doing

When humans produce a common response to an array of distinct objects or series of events, we can infer that they share an equivalence or category identity. People can effortlessly use tools interchangeably, which possess similar functionalities or attributes that allow them to solve the problem at hand, even with little learning or practice and often involving an act of invention.

Here, the aim is to train the robot to identify that a set of utensils generate the same response based on one or more mutual attributes and can, therefore, be used coequally, without further learning. The core activity is to pour liquid into a cup from a utensil. While it is interesting how fundamental concepts are formed (e.g., bottle), this is not central to the aim, which is to favourably “replace” the object in the event context with an equivalent exemplar pertaining to yet a higher-level joint constraint, that of being a container. Hence, the discriminable features of the individual objects are not accounted for (e.g., telling apart a bottle from a carafe). The ability of the robot to identify the utensils uniquely arises from binding the visual stimuli (object recognition) with linguistic input (label dictated by the human tutor), to generate the <form, form> mappings explained in the method of binding domain-specific information (section 4.3.2). The integrated vision component of the virtual PR2 robot tracks the objects and produces a stimulus that is processed by the learning framework as a symbolic representation. The anchor between the object and its label is uniquely and permanently created, through the association symbolic capacity modelled in Chapter 4. This is an overly simplified modelling of concept formation (Bruner & Austin, 1986). These concepts are later retrieved in the behavioural experiments as part of more convoluted categories and event constraints that involve both observable and lexicalised attributes.

The data are collected as follows. Initially, the long-term memory (LTM) is populated with the objects (as in 4.4.1.2) and action primitives (as in 4.4.2.2) that will be used in the experiments. The objects are exemplars and close-non exemplars (similar attributes) of the involved categories of [container] and [liquid], whereas action primitives are elementary operations pre-determined on the virtual PR2 robot. Their <form, form> association is retained in the LTM. The LTM maintains also lexicalised categorical attributes in the form of declarative phrases dictated verbally by the human tutor (multimodal information). Table 5 illustrates only the verbal explicit knowledge.

Table 5 The verbal explicit knowledge maintained in the long-term memory (LTM) of the cognitive architecture. It comprises information about objects that belong to the category of containers (exemplars) and of objects that do not
(non-exemplars). Non-exemplar objects possess close-in attributes with the containers, e.g., "the hat-trick has an opening" is similar to "the jug has an opening" (column 2). Moreover, the robot knows a range of liquids, which are discriminable based on their colour attribute.

<table>
<thead>
<tr>
<th>Object</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exemplars of containers</strong></td>
<td></td>
</tr>
<tr>
<td>bottle</td>
<td>the bottle has a neck</td>
</tr>
<tr>
<td></td>
<td>the bottle has a screw-cap</td>
</tr>
<tr>
<td></td>
<td>the bottle is made of plastic</td>
</tr>
<tr>
<td></td>
<td>the bottle is a container</td>
</tr>
<tr>
<td>jug</td>
<td>the jug has an opening</td>
</tr>
<tr>
<td></td>
<td>the jug has a handle</td>
</tr>
<tr>
<td></td>
<td>the jug is made of glass</td>
</tr>
<tr>
<td></td>
<td>the jug is a container</td>
</tr>
<tr>
<td>cup</td>
<td>the cup has an opening</td>
</tr>
<tr>
<td></td>
<td>the cup does not have a handle</td>
</tr>
<tr>
<td></td>
<td>the cup is made of glass</td>
</tr>
<tr>
<td></td>
<td>the cup is a container</td>
</tr>
<tr>
<td><strong>Non-exemplars of containers</strong></td>
<td></td>
</tr>
<tr>
<td>hat-trick</td>
<td>the hat-trick has an opening</td>
</tr>
<tr>
<td></td>
<td>the hat-trick is large</td>
</tr>
<tr>
<td></td>
<td>the hat-trick is made of textile</td>
</tr>
<tr>
<td>cylinder</td>
<td>the test-tube has an opening</td>
</tr>
<tr>
<td></td>
<td>the test-tube does not have a handle</td>
</tr>
<tr>
<td></td>
<td>the test-tube is cylindrical</td>
</tr>
<tr>
<td></td>
<td>the test-tube is made of glass</td>
</tr>
<tr>
<td>vase</td>
<td>the flower-pot has an opening</td>
</tr>
<tr>
<td></td>
<td>the flower-pot does not have a handle</td>
</tr>
<tr>
<td></td>
<td>the flower-pot is made of plastic</td>
</tr>
<tr>
<td><strong>Exemplars of liquids</strong></td>
<td></td>
</tr>
<tr>
<td>water</td>
<td>the water is liquid</td>
</tr>
<tr>
<td></td>
<td>the water is transparent</td>
</tr>
<tr>
<td>milk</td>
<td>the milk is liquid</td>
</tr>
<tr>
<td></td>
<td>the milk is white</td>
</tr>
<tr>
<td>wine</td>
<td>the wine is liquid</td>
</tr>
<tr>
<td></td>
<td>the wine is red</td>
</tr>
<tr>
<td>juice</td>
<td>the juice is liquid</td>
</tr>
<tr>
<td></td>
<td>the juice is orange</td>
</tr>
<tr>
<td>beer</td>
<td>the beer is liquid</td>
</tr>
<tr>
<td></td>
<td>the beer is amber</td>
</tr>
<tr>
<td>coffee</td>
<td>the coffee is liquid</td>
</tr>
<tr>
<td></td>
<td>the coffee is brown</td>
</tr>
</tbody>
</table>

5.3.1.1. From language to behaviour: step-by-step

The experience of the robot to manipulate objects is built in a series of runs of the snowballing artifice (see 4.3.2.1). The test toolchain is constructed gradually at runtime starting from action primitives to higher-level motor exploration (Figure 25). The robot is initially trained to execute simple lower-order actions using a one-occasion sample with the object bottle. The process is then repeated during runtime testing with other novel objects, and the experiences are self-memorised in the robot’s long-term memory. The runtime test is used to validate the ability of the robot to convert (novel) high-order language instructions into symbolic representations of robot actions. It is modelled as a natural communication between the human and the robot, verbally guiding the robot to interact with the objects in the workspace, using natural language (NL) instructions of increasing level of abstraction (Figure 25).
5.3.1.2. First higher-order learning

Initially, the robot is probed to discern containers from non-containers, by training the instruction, *take the container*. Notice how the robot has been taught to execute simple actions one step at a time on each object individually, however, the instruction is formulated on the higher-level category [container]. Therefore, the robot must first identify the object in the workspace (*concept formation*) and fathom whether it belongs to the category of containers (*concept attainment*), by recalling this attribute from long-term memory. Next, it will attempt to execute the motor actions that are specific to the observed object (one occasion learning, Figure 25).

The data are collected as follows: The robot hears the spoken instruction and observes the object in front (Figure 26). By means of object recognition, the visual stimulus is encoded to symbolic representation. Both the auditory (linguistic) and visual signal (pseudo-sensory representation) are processed by the working memory of the learning framework. While building the mental action sequence, the framework will store the event constraint as a goal to be readily accessible for decision-making. Should the constraint not be met, i.e., *the observed object is not a container*, the learning framework will cease the mental action sequence and no output will be produced. Therefore, one occasion learning example is sufficient to discern cases of exemplars from non-
exemplars in the context (to abstract from their differences and similarities) and devise the voluntary control competence to proceed with or cease the task autonomously.

After the initial learning, the robot is probed with other objects that are either exemplars of the category [container] or not. The robot retrieves the mental action sequence that was created during training to build the output preferentially following the same line of reasoning. In the case of exemplars of containers, it will produce a valid output, whereas, for non-exemplars, the mental sequence will be interrupted for failing to verify the event-imposed attribute (i.e., object is not a container) (Figure 26). The robot self-rejects these cases, as inappropriate for the context, and as such, verifying criterion 1 of concept understanding (C1).

Figure 26 Discerning examples from non-examples of [container](s). The robot is trained with a single example to take the container it sees in front of it. The spoken instruction “take the container” is combined with the percept data received by the virtual robot in the Webots world. The learning framework that elaborates the input (linguistic + non-linguistic), uses its knowledge from long-term memory to validate if the perceived object possesses the attribute of container. In the test, the robot must discern the objects appropriately, based on this property. Non-examples of containers share close-in attributes with the containers; however, these attributes are non-critical in the context. The robot is tasked with retrieving only relevant information (critical attribute) that responds to the human’s request.

5.3.1.3. Second higher-order learning

In the event of pouring a specific liquid to a cup, further constraints are imposed besides inferring which utensils are adequate. The robot is trained to obtain the ability of discerning liquids given their colour identity using the instruction, add <liquid> to the cup, where <liquid> is any of the six illustrated in Table 5. Colour was set as a defining attribute and the exemplars were selected with clear-cut boundaries (i.e., of one and a different colour), to reduce the cognitive load of classification. Realistically, one would require deeper reasoning to separate two liquids in their meaningful category in the physical locale, for example having to choose between hot chocolate and coffee (brown) by observation or taste.

The virtual robot is situated in front of a table, in which there is a cup and a random utensil (Figure 27). For simplicity, the colour of the liquid is assumed to be the colour of the container, given the slight limitations of the virtual simulator. Other parameters are not considered, e.g., whether the utensil is full or empty. The robot’s vision is used to recognise the objects and their colour feature
that are encoded to symbolic representations and permanently mapped to the respective lexicalised terms (*concept formation*) to be processed by the working memory.

![Training of the concept-learning task.](image)

*Figure 27 Training of the concept-learning task. The robot is trained with a single example to add water to the cup from a bottle. For visualisation, the blue colour in the figure represents water (transparent). The speech data (add water to the cup) is combined with the percept data from the observed scene (utensil: bottle, colour: transparent, target object: cup). The latter are represented conventionally through linguistic symbols. The robot verifies two constraints: a) is the utensil a container, b) is the liquid observed in the scene water (by colour assessment). Should these constraints be met, the robot uses the snowballing toolchain to retrieve the earlier-learned primitive actions to complete the task. Otherwise, it ceases the task. In the test, the task is repeated consistently by replacing the bottle with other containers (jug, cup) holding liquids of random colour.*

The robot was trained on the one occasion example of pouring water into a cup from a bottle. It will attempt to verify the event-imposed constraints of procuring an [container] of [transparent] colour (i.e., water) and self-regulate the course of action from the snowballing extrapolation as in *Figure 25*. When a higher-order instruction is added to the toolchain (*add*), the exploration experience of procuring the bottle is retrieved impromptu and fused with further motor modalities corresponding to the higher-order manipulation verb. This experience does not require training *de novo*. Which primitive action (step 3) is added to the toolchain is self-inferred by the robot, given the linguistic stimuli offered by the human when guiding the task build-up (e.g., to add liquid, step 3 pour the liquid). This learning artifice allows higher-order modalities to evolve from preceding tasks or earlier-acquired low-level experiences, by adding more convoluted steps in the toolchain and, thus, expanding the robot’s behavioural repertoire.

### 5.3.2. Scenario 1: Concept attainment of increasing abstraction

Humans demonstrate understanding when deploying their acquired knowledge to solve problems in unfamiliar contexts, with little to no learning (2). To probe this high-level cognitive skill in the robot, the concepts learned in the foundation task are situated in two higher-complexity scenarios. The first scenario investigates the role of linguistic stimuli in the autonomous decision-making of the robot, in how it affects the categorisation, abstraction and voluntary control of the robot in the new context.

91
The experiment is devised as follows. The virtual robot is situated in front of a table with three objects, a cup, a teabag, and a bottle of water (Figure 28). The percept representations of the objects are extracted from the robot’s vision and mapped to their respective forms in permanently memorised pairs, by binding information from multiple domains as explained in section 4.3.2. The primitive actions to manipulate the objects are the same as those trained in the foundation activity (grasp, lift, throw, etc.). Through a runtime snowballing, the human uses simple natural language commands to initially guide the robotic model to use the novel tools specific to the scenario (cup, teabag) – learning snowballing. The robot gradually self-converts language stimuli into motor exploration and generates higher-order motor modalities from the interaction of action primitives with linguistic experience. At the end of the runtime test, the robot has learned to use a cup by grasping, lifting, and dropping it on the table (\textit{cup\_grasp\_lift\_table\_drop}) and the steps to put a teabag inside the cup (\textit{teabag\_grab\_pickup\_cup\_throw}). Instead, the act of using a bottle to pour water is retained in long-term memory as an earlier-learned experience from the foundation concept-learning activity, i.e., not re-trained.

The robot was appropriately trained on one occasion to prepare tea using a \textit{bottle of water} (Figure 28). No observation or demonstration of the action was introduced to facilitate the task, i.e., the innate steps of preparing tea from a \textit{clean slate} to replicate the behaviour in future endeavours. Instead, the robot was only verbally dictated the lexicalised property \textit{to make tea you need water} as a factual affirmation. This is utilised to make the requisite inferences on the instances that are appropriate for the context. In other words, the selection range of the available liquids and colours is narrowed down to \textit{water} and \textit{transparent} whilst all other candidate members of their respective categories must be rejected. The scenario is trained by uttering the instruction “make me tea”. This instruction hints not neither the liquid nor the utensil (e.g., “\textit{make me tea with the bottle}” or “\textit{make me tea using water}”). This information is retrieved from the virtual PR2 robot directly from the scene (Figure 28), i.e., the robot observes a \textit{bottle} of \textit{transparent} colour in the workspace and feeds this back to the learning framework in symbolic representations.

![Figure 28](image.png)

**Figure 28** Scenario 1: Concept attainment of increasing abstraction. The robot is trained with a single example to prepare tea using a bottle, given the verbally dictated lexicalised attribute “to make tea you need water” (LTM). The previous experience of adding water to the cup is not re-trained but retrieved directly from past experience with the concepts from the foundation concept-learning task, which are memorised permanently in long-term memory. In the test, scenario 1 is repeated consistently, using other utensils (jug, cup) with different liquids. The robot must colour-discriminate liquids and execute the task only with exemplars of water containers.
The event-imposed constraints are verified similarly as in the foundation concept-learning activity. When the robot infers that the involved concepts are germane in the context (the observed liquid matches the lexicalised attribute *to make tea you need water*), it binds together [bottle] and [water] to retrieve from long-term memory the earlier learned experience of pouring water to the cup, without re-training this praxis. Hereby, preceding tasks are successfully transferred to new higher-complexity scenarios, given the skill to make non-trivial generalisations from lexicalised attributes on the learned concepts, and as such emphasising the role of language in *concept attainment*. To select the concepts that fit this context, the robot ought to abstract the attributes of [water] from its similarities with other instances (attribute: liquids - adequate in a preceding task but not in the current event) and their differences (tea can be made only with water – language). It is imperative to highlight that the robot has never been taught to cease a task when the constraints are not met (e.g., if it is given a bottle of wine). This competence is autonomously generated by the robotic model by virtue of assessing the conceptual content of the event.

### 5.3.3. Scenario 2: Conceptual content in polyglot contextualisation

A certain concept can be labelled by two or more synonyms in the same language or by different terms in distinct languages. These terms are linguistic forms in which humans express in words the meaning of what they see or sense through their bodies. The concept meaning can, therefore, be used as a pivot or common ground representation between the different terms/forms. When a label is attached to a concept, it becomes part of its representation and it can trigger (auditory stimuli) the same conceptual content as other perceptual stimulations of the concept (somaesthetic, visual, olfactory and taste).

The motivation to utilise two distinct human languages in scenario 2 is twofold:

1. Intuitively, the concurrence between phrases expressed in different languages is much more insignificant than that between phrases of the same language, in which one or more concepts are substituted with a synonym. Hence, it can be assessed with higher certainty whether the robot is identifying concepts correctly or relying on phrase similarity to build a valid output.

2. Two human languages represent separate knowledge sources, whose combination enables the robot to expand its semantic and behavioural repertoire, e.g., in partially incomplete scenes. The artifice of exploiting multiple knowledge sources is a fine practice in dialogue agent configuration (Jiang et al., 2015). The skill of exploiting all available information is central to task-solving. In the method described in this chapter, the knowledge implicated in each language source (English/Italian) is tailored as complementary information on the involved concepts (i.e., not duplicated). To solve the task, the robot must put together cues and pieces of information from both sources, which make the complete puzzle of requisite attributes and constraints in the context. As such it can validate if the linguistic information (concept labels in each language) is sufficient to retrieve the representation of the concept in a locale where this representation and the experiences around it are not directly learned (see 5.3.3.3). The emergence of conceptual content from multilingual domains can lead to robotic agents that operate efficiently in multilingual environments or are transferred with ease from one environment to the other with little learning.

#### 5.3.3.1. Concepts as pivots

In the word-learning stage, infants self-map the objects they see with the words they hear to describe them in a pragmatic frame (Markman, 1989; Bruner, 1985). Compound bilinguals acquire
at least two linguistic terms for the same observed entity, even before formally learning the languages (Cook, 1992; Crinion et al., 2006). Similarly, the method proposed here utilises strongly observable concepts whose (representation) meaning is generally omnipresent and intentionally leaves out aspects of localisation or socio-cultural-dependent meaning of a concept (Wierzbicka, 1992). A certain concept is expressed through different terms in different languages. The technique of binding information from linguistic and non-linguistic domains, explained in section 4.3.2 is adopted to create <meaning, form> (<_form, form>) pairs in each language. The meaning of a concept is associated with the variable language-dependent form(s), in which the (same) concept is formulated, thus the concept becomes a pivot between the semantic labels. This is illustrated in Figure 29.

A major difference between the paired representations with the standard method (section 4.3.2), is that, here, an auxiliary language indicator (_en, _it) is added to the pairs. In analogy with the language-sensitivity of the bilingual brain, which enables bilinguals to discern and control the language they hear and use (Crinion et al., 2006), applying a language indicator in the <_form, form> pairs, increases the framework’s awareness and its ability to disambiguate the language in which the concept is expressed. Hereby, it can self-transpose from one knowledge source to the other during the decision-making process. The language indicator is extracted via automatic language recognition and elaborated in the learning framework when generating the respective symbolic associations between concepts and their internal representations.

Using two natural languages as knowledge sources not only intuitively decreases the concurrence between words and phrases of the same language but also spikes the number and complexity of mental actions that the working memory will undertake to build a valid output.

Let us consider the following example.

The chocolate is food.
The chocolate is sweet.

*Human:* tell me a sweet food
To respond to the instruction, the working memory will retain the keywords *sweet* and *food* in the focus of attention and goal stack, to retrieve the closest relevant phrase and answer *chocolate*. It is evident that the above phrases share at least one word and/or grammatical construction (e.g., sentence order/structure).

Next, let the same phrases be given in distinct languages.

```
The chocolate is food.
Il cioccolato e’ dolce.
Human: tell me a sweet food
```

Albeit the meaning is not altered, it is burdensome or unattainable for the working memory to build a valid output for the instruction, as there are no joint or concurrent word cues. Hence, further cross-linguistic references are required. Let them be:

```
_chocolate _en chocolate
_chocolate _it cioccolato
_sweet _en sweet
_sweet _it dolce
```

where *chocolate* is the visual (and/or taste stimulus) representation of the chocolate and *sweet* is the taste stimulus representation of something tasting sweet. The above associations indicate that the robot knows the linguistic label of the concepts in both languages (retained in memory).

To produce the same answer *chocolate*, the working memory here must elaborate four more phrases to obtain the cross-lingual cues that are needed to build the output. The referent *food* is pushed into the goal stack. The WM first retains the word *sweet* and uses the pivot representation to extract the equivalent label *dolce* to retrieve the phrase *il cioccolato e’ dolce* from long-term memory. The representation *chocolate* is then extracted similarly, to obtain the equivalent term *chocolate* and used as a cue, along with the word *food* stored as a goal, to retrieve the relevant phrase *the chocolate is food* and send *chocolate* to the output.

### 5.3.3.2. Behavioural emergence in multilingual domains: scene 1 (English)

The task of scenario 2 involves the same robot behaviour trained and memorised during the foundation concept-learning activity. The expected outcome is to add a certain liquid to a cup from a utensil. The setting of scenario 2 is similar to the foundation task: the virtual robot is situated in front of a table, in which there is a bottle and a cup. In addition, there is a human actor in the scene. A central difference with the concept-learning task is that the enacted experience in scenario 2 is not re-trained but retrieved from long-term memory as an earlier-learned experience, under novel constraints in the running context, i.e., the concept-learning task is a retained experience readily accessible to be recalled in a new context. These constraints, which impose the robot to make a reasonable decision during task-solving, are the involved actor’s beverage preferences that are added to the scene. Such attributes or constraints are lexicalised and not observable (our likes or dislikes are generally abstract notions), hence, the aim is to assess in what way and to what degree of skill will the robot use these attributes to discern exemplars from non-exemplars in the given event.

With that in mind, scenario 2 resembles scenario 1. The staple contrast is that the lexicalised attributes/constraints (the actors’ potable preferences) are dictated in simple phrases in the *Italian* language. This scenario assumes that the robot has a limited linguistic experience in Italian and a
Chapter 5: High-Level Cognitive Modelling in Robots

basic symbolic capacity to associate labels with representations, trained only on a restricted number of concepts, using the domain-specific information binding methodology. This long-term knowledge is illustrated in Table 6. The long-term memory is populated with more affirmative phrases and multimodal representations that relate to the actors’ food preferences to validate the generalisation and abstraction competence of the robot on knowledge of similar context that does not fit the event constraints.

Table 6 The knowledge representation in long-term memory (LTM). The robot has memorised the cross-lingual pivot representations from preliminary training (column 3). These associations of different semantics co-exist in memory and the explicit knowledge from different languages is intertwined. The robot self-retrieves the relevant lexical term using the language indicator ( _en, _it). Notice how there are no direct representations of central concepts used here in the Italian language (e.g., bottle, jug,... or colours: transparent, red...). The event constraints (the actor’s beverage preferences) are made known to the robot in the form of declarative phrases in the Italian language. Auxiliary phrases on the food preferences are introduced to assess with certainty the robot’s generalisation capacities during decision-making. *The translation of the Italian phrases is only shown to improve the reader’s comprehension but are not part of the information stored in the LTM of the learning framework.

<table>
<thead>
<tr>
<th>Beverage preferences</th>
<th>Food preferences</th>
<th>Cross-lingual pivots</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Mary piace il vino.</td>
<td>A Mary piace la pizza.</td>
<td>_drink _en drink</td>
</tr>
<tr>
<td>A John piace il latte.</td>
<td>A John piace la pasta.</td>
<td>_wine _en wine</td>
</tr>
<tr>
<td>A Lucy piace il succo.</td>
<td>A Lucy piace la torta.</td>
<td>_milk _en milk</td>
</tr>
<tr>
<td>A Peter piace il caffè.</td>
<td>A Peter piace la frutta.</td>
<td>_juice _en juice</td>
</tr>
</tbody>
</table>

English Translation (*)

Mary likes wine.
John likes milk.
Lucy likes juice.
Peter likes coffee.

Mary likes pizza.
John likes pasta.
Lucy likes cake.
Peter likes fruit.

Other declarative phrases

Wine is a drink.
Milk is a drink.
Juice is a drink.
Coffee is a drink.

* The translation of these phrases is only shown to improve the table comprehension.

The robot was trained on one occasion to serve a drink to the actor Mary from a bottle (Figure 30), given the lexicalised property A Mary piace il vino (see Table 6 for meaning). Note that the uttered instruction and the verbal affirmation are of distinct languages. The robot must be able to identify the concepts that are jointly involved in both phrases and their associated attributes to retrieve the content. The working memory builds the mental action sequence by initially validating what to serve [drink] to the referent actor. By means of such reasoning, the food preferences of the involved actors are never retrieved as relevant information, because the instruction involves the concept of [drink], which is retained and associated with the information stored in memory related to [drink]. After inferring the intent, the robot must verify the anticipated identity of the concepts in the workspace. The constraints of the event (a bottle, red liquid) are verified similarly as in the foundation concept-learning task and/or scenario 1. The robot then self-retrieves its earlier-learned
experience with the concepts to add wine to the cup from the bottle (as in the concept-learning task and maintained in long-term memory). When converging the physical manifestation (meaning) of the manipulation verb serve, the snowballing toolchain is recalled to indirectly self-link the constituting actions and relevant tools in the context, i.e., the meaning of to serve in this context will be to pour a drink to a cup, using an available utensil (_bottle _grasp _lift _cup _pour). In the robot’s motor map, the motor exploration of the verb serve will be the combined trajectories of the action primitives grasp, lift, pour that are enacted on the objects [bottle] (origin target) and [cup] (destination target).

A compelling situation of this scenario is, for instance, how a multilingual-speaking robotic model that receives stimuli from people speaking different tongues, can reuse its grounded experience impromptu to respond to new linguistic information at a satisfactory degree of skill without retraining that experience to accommodate the new semantics, by means of the attained conceptual content. The linguistic input alone triggers the same mental states as the sensorimotor experience to reproduce that behaviour directly.

**Figure 30** Scenario 2: Conceptual content in polyglot contextualisation. The robot is trained with one example (one per language) to serve to some person a drink of their preference. The scenario unfolds from the concept-learning task, with the novel lexicalised constraint of discerning liquids based on their colour AND the lexicalised attribute of “preferred drink”. That means previous experiences trained and memorised in the concept-learning task are retrieved directly to execute the task upon verifying the requisite conditions. In the test, the robot-model is consistently tasked with serving a drink to a certain actor: which drink is self-determined by the robot based on the lexicalised constraint of preference and its identity in the scene is self-verified via colour observation. The robot demonstrates identical behaviour in each of the languages, albeit having very limited experience in Italian (no snowballing), using concepts as common ground to retrieve the learned experiences.

5.3.3.3. **Novel contextualisation of earlier-learned behaviour: scene 2 (Italian)**

The capacity of our human brain to adapt to diverse environments, be those physical habitats or socio-cultural contexts, is largely devoted to our cognitive ability (Boyd et al., 2011). The method described in this passage attempts to demonstrate how robots could potentially achieve such competence by appropriately developing and understanding concepts. It seeks to verify whether a certain accumulated experience can be successfully reused directly when contextualised in a novel semantic scene. Therefore, the robot was probed to replicate the same behavioural praxis as in 5.3.3.2, this time queried directly in Italian. To succeed, the robot must analyse the event in terms of the concepts surrounding it. From a computational angle, this challenges the learning
framework’s capabilities on a high hardware resource-demanding task, in turn, allowing to assess its robustness.

The setting of this scenario is identical to Figure 30, with the main difference being that the user utters the instruction directly in Italian (servi una bibita a Mary). The symbolic associations that the robot has retained in memory in Italian are given in Table 6. Notice how there are no direct representations of central concepts used here in the Italian language (e.g., for bottle, jug, cup or for colours: transparent, red...). Moreover, the robot has never been trained any experience (linguistic or not) in the Italian language, either in the concept-learning task or scenario 1 (i.e., the learning snowballing). On this note, the robot is challenged to fathom the intent of the instruction by identifying the involved concepts and, next, self-reproduce the desired outcome. This capacity is facilitated by the linguistic labels associated with the representations of the concepts: while labels differ in language, the grounded conceptual representations pivot to those labels (Figure 29). The real-world representation of the concepts is retrieved from the locale of the robot (e.g., bottle, red liquid). Hence, the anticipated identity of the liquid is validated as in former experiences and is independent of the language. For instance, if the robot observes a bottle in the workspace, it needs not know explicitly that the bottle in Italian is called bottiglia, to use the bottle as a container (experiential conceptual content). Similarly, knowing the explicit colour labels in Italian is not central to the intelligent understanding of how to discriminate liquids by colour. Such experiences are therefore retrieved directly with no cost of re-learning. The WM is only trained to use the labels associated with the concept representations, to assess the critical observable and lexicalised attributes of the concepts scattered across languages and faithfully reproduce the act of serving a (specific) drink, without any previous direct experience in that locale.

5.4 Results

This section illustrates and delves into the obtained results of the experiments in the attempt to validate the three criteria of understanding: identifying a concept by its salient attributes over its variable non-defining attributes (C1) and habitually discriminating exemplars from significantly close non-exemplars (C2) on unlearned endeavours with the concepts (C3).

5.4.1 Foundation concept-learning task: learning by doing

In the test, the robot is probed using objects that meet the constraints (A, D) or not (B, E) (Figure 31). To generalise, the robot must execute the task consistently each time the conditions are met, on novel entities from groups A-E not used in the learning example. This is illustrated in Table 7.
“the hat-trick has an opening” resembles “the jug has an opening”). This knowledge is stored in long-term memory. The first event constraint is to select only examples of containers (group A) and reject all non-examples (group B). The second constraint is to separate liquids based on their discriminable colour. An example of the concept refers to the queried liquid (speech) matching its observed colour identity in the scene (group D); otherwise, it refers to non-examples (group E).

A total of 14 cases, in which the event constraints were met (A, D) and 60 cases, in which at least one was not (B, E) were investigated. The robot could successfully identify all examples of the concept of [container] (C1), select liquid exemplars correctly to pursue the task [14] while rejecting all [60] cases of non-examples (C2), and extend these conceptual attributes across the other members of their respective categories (C3), without a priori learning. For instance, the robot had never learned to use a jug when the task was trained. Instead, it has been taught that, given that bottle is a container, it is germane for the task. The robot is then able to extend this property autonomously across other candidate members (e.g., the jug). Similarly, it applies the colour-discrimination constraint to all other liquid types (C2) not presented during learning (C3). What is more, having learned that water is liquid, and that liquid is poured, it inferred that all liquids can be poured, self-producing an exact behaviour with all novel instances of group D, as if they were the same thing (i.e., abstraction), from one learning sample alone (C3). Using the verbal stimuli “is a container”, “is a <colour>”, the robot can classify the instances pragmatically into the same category and produce similar behaviours, which are different from the instances that do not share the same attribute (different category - not a container, not the right colour) (categorisation). The robot could self-regulate its decisions to proceed with or cease the task, by means of reasoning on the critical constraints of the context (voluntary control).

Table 7 The one-occasion learning sample and generalisation sets used in the foundation task. Test case 1 includes cases of category members with similar criterial attributes, which meet the event constraints. Test case 2 includes cases in which the colour constraint of the liquid is not met. The robot was able to execute the task successfully for all test case 1 and cease the task otherwise (test case 2).

<table>
<thead>
<tr>
<th>TRAINING SET (one-occasion)</th>
<th>Event attributes</th>
<th>No. of learning examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Container</td>
<td>Instructed liquid</td>
<td>Observed liquid colour</td>
</tr>
<tr>
<td>[Image of bottle]</td>
<td>Water</td>
<td>transparent</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TEST SET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test case 1: Exemplars of containers, exemplars of liquids</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Event attributes</th>
<th>No. of testing examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Container</td>
<td>Instructed liquid</td>
</tr>
<tr>
<td>[Image of cup]</td>
<td>Water</td>
</tr>
<tr>
<td>[Image of jug]</td>
<td>Wine</td>
</tr>
<tr>
<td>[Image of bottle]</td>
<td>Milk</td>
</tr>
</tbody>
</table>
Test case 2: Exemplars of containers, non-exemplars of liquids

<table>
<thead>
<tr>
<th>Container</th>
<th>Instructed liquid</th>
<th>Observed liquid colour</th>
<th>No. of testing examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>red white orange amber</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>Wine</td>
<td>transp. white orange amber</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>Milk</td>
<td>red transp. orange amber</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>Juice</td>
<td>red white transp. amber</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td>Beer</td>
<td>red white orange transp.</td>
<td></td>
<td>12</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td></td>
<td><strong>60</strong></td>
</tr>
</tbody>
</table>

Total no. of testing examples 74

The behaviour of the virtual PR2 robot in the concept-learning task is demonstrated in Video 4.

5.4.2 Scenario 1: Concept attainment of increasing abstraction

The lexicalised affirmation (conceptual attribute) *to make tea you need water* allows the robot to make nontrivial inferences on which instances are adequate in the current context (*C1, C2*). A central aspect of the obtained results for this scenario is that the robot directly reuses the experience learned in the foundation task, which is permanently retained in long-term memory. Re-using earlier-learned experiences impromptu is a desideratum for cognitive robotics research.

The test was used to assess the robot’s competence to prepare tea with any water container (group A) and reject examples of other beverage containers (group B) (*Figure 32*). The robot was consistently probed with random instances (*Table 8*). In all 14 studied cases, the robot could successfully discriminate acceptable liquid exemplars (group A) from non-exemplars (group B) in the event context (*C2*), by means of autonomous inference on the linguistic constraint. Moreover, it could transfer the experiences acquired in the foundation task to a peculiar context, appropriately extending those behavioural praxes across all novel instances of the same category (group A) without pre/re-training (*C3*). In response to the linguistic stimulus, the robot could change its initial
categorisation decisions (i.e., objects that are containers belong to the same category) to produce new categories (i.e., containers of water are of a different category from other potable containers) (abstraction) and, as a direct consequence, new behaviours accordingly (voluntary control).

Figure 32 Scenario 1: Concept attainment of increasing abstraction. The robot is trained to make tea using a bottle of water. Learning to use the involved tools (cup, teabag) is achieved via the snowballing artifice. Group A includes instances that can be used to prepare tea, whereas group B involves instances that cannot. The linguistic input alone can successfully narrow down the available concepts into those that fit the running context, by abstracting from their differences and similarities (C1, C2) in new encounters not anticipated at learning (C3).

Table 8 The one-occasion learning sample and generalisation sets used in scenario 1. Test case 1 includes examples in which the involved concepts possess attributes that meet the event-imposed constraints. Test case 2 includes cases of non-exemplars. The robot successfully executed the task in all test cases 1 or ceased the task otherwise (case 2).

<table>
<thead>
<tr>
<th>TRAINING SET (one-occasion)</th>
<th>Event attributes</th>
<th>No. of learning examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Container</td>
<td>Observed liquid colour</td>
<td>Container transparent 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TOTAL 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TEST SET</th>
<th>Event attributes</th>
<th>No. of testing examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test case 1: Exemplars of containers, exemplars of water</td>
<td>Event attributes</td>
<td>No. of testing examples</td>
</tr>
<tr>
<td>Container</td>
<td>Observed liquid colour</td>
<td>Container transparent 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TOTAL 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test case 2: Exemplars of containers, non-exemplars of water</th>
<th>Event attributes</th>
<th>No. of testing examples</th>
</tr>
</thead>
</table>
The concept-like behaviour of the virtual PR2 robot in scenario 1 is demonstrated in Video 5.

5.4.3 Scenario 2: Conceptual content in polyglot contextualisation

Humans can understand concepts expressed in distinct verbal expressions. “I like milk” and “mi piace il latte” (Italian) are distinct utterances, whose predicates are different formulations of the same concepts. The experiments devised here seek to validate if the robotic model can emerge similar competence.

Figure 33 illustrates the robot waiter tasked with serving a drink to four people (including Mary), probed with utensils of a drink of preference (group A) or not (group B), in either language.

<table>
<thead>
<tr>
<th></th>
<th>red</th>
<th>white</th>
<th>orange</th>
<th>amber</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>

Total no. of testing examples 14

Figure 33 Scenario 2: Conceptual content in polyglot contextualisation. The robot is tasked with serving a preferred potable to a person (up left). The preferences are formulated in simple Italian phrases (upright). Further food preferences are dictated to assess the ability of the robot to discriminate the concept of [drink]. Group A includes examples of concept appropriateness, in which the event-imposed constraints are met. Group B includes cases in which the observed drink does not match the anticipated identity of the actor’s beverage preference. The task is repeated two times: (a) robot queried in English and (b) robot queried in Italian.

Scene 1 (English): A total of 11 cases in which the constraints of the event were met, and 36 opposing cases were investigated (Table 9, Learning Examples English). The robot would only procure the object should it be a container (self-inference, CI) and should the observed potable
match the anticipated identity of the actor’s preference (C2), while successfully rejecting all (36) non-exemplars of either concept. It habitually executed the task with novel members of the respective categories that possessed salient attributes appropriate for the context (C3), learning only from one occasion. It changed its categorisation decision in response to the verbal stimulus and produced new behaviour accordingly. Reminiscent of scenario 1, past experiences are not re-trained, but self-retrieved from the long-term memory.

Scene 2 (Italian): Further 11 cases of concept exemplars and 36 non-exemplars were explored (Table 9, Learning Examples Italian). The increased complexity implication of the working memory to achieve cross-linguistic information retrieval when queried in Italian is explained in section 5.3.3.1). Even under limited linguistic and motor experience and the spiralling neural processes required to generalise across multilingual domains, the robot manifested an exact behaviour as in the English counterpart, demonstrating all three criteria of understanding in the entire test set. Having no past experiences in an Italian context, the robot could reuse those learned in English in a wholly new semantic encounter (Italian) without re-training (C3), after having successfully self-assessed the critical attributes of the involved concepts and having identified the goal of the task. For instance, when a robot had conceptually determined that [Mary] and a [red liquid] are in the scene (concept assessment), it needed not re-learn how to fill a cup with wine (task goal). This might drastically reduce the need to train robots de novo in each natural language, but rather exploit conceptual content as a common ground, to transfer the behavioural repertoire faster in new linguistic scenes.

Table 9 The one-occasion learning sample and generalisation sets used in scenario 2, In English and Italian. Test case 1 refers to examples of concepts that meet the constraints in the context. Test 2 involves cases in which these constraints are not met. In both linguistic scenes, the robot was able to solve all examples of test case 1 and cease the task appropriately in test case 2, by means of self-inference on the critical constraints.
Chapter 5: High-Level Cognitive Modelling in Robots

### Test case 2: Exemplars of containers, non-exemplars of beverages

<table>
<thead>
<tr>
<th>Container</th>
<th>Actor</th>
<th>Preferred beverage</th>
<th>Observed colour</th>
<th>Learning examples (English)</th>
<th>Learning examples (Italian)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>wine</td>
<td>white</td>
<td>orange</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>John</td>
<td>milk</td>
<td>red</td>
<td>orange</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Lucy</td>
<td>juice</td>
<td>red</td>
<td>white</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Peter</td>
<td>coffee</td>
<td>red</td>
<td>white</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

**TOTAL** 72

**TOTAL** 22

**Total no. of testing examples** 94

The concept-like behaviour of the virtual PR2 robot in scenario 2 is demonstrated in Video 6.

5.5 Conclusion

This chapter dwelled on two focal hypotheses on humanlike conceptual development: Can an artificial cognitive model develop appropriate conceptual content and structure its knowledge representation (i.e., organise its experience into coherent patterns) to draw non-trivial inferences on its locale? What is the cognitive role of language in this conceptual development? When can we assume that such models understand the concepts they use?

The chapter describes three experiments with a simulated humanoid robot that seek to study central language-facilitated cognitive abilities, such as the ability to act (perceive, manipulate) in a real ambience where intelligent behaviour is required and in which the emergence of concepts can be verified. A notable feature of the proposed method is the competence of the robot for continuous and impromptu learning of new practices in/from its physical locale in task-accomplishment goals using (human) language as the natural communication interface to emerge its conceptual content. The methodology offers significant refinements to the procedural mechanisms of the learning framework leveraged here to allow apt learning of the robot from one occasion, which is the natural way humans learn and a much-desired goal for the next-generation AI models.

The two higher-complexity scenarios seek to investigate the critical role that language plays in the process of decision-making. Affirmations (“you need water to make tea”), in particular subjective forms of expression (“I prefer this”), allow many forms of reasoning that are rather arduous without concepts. Adapting a learning procedure that resembles how our working memory combines pieces
of information into a meaningful outcome, underpins much of the competencies demonstrated in the experiments described here. More compelling is how the robot decides autonomously whether to pursue or cease a task, though not taught to do so specifically. Instead, it has learned to “reason” on the significant attributes that take place in the context, by recalling what it knows (memory) in what it is (current scene). On one hand, this allows the robot to produce similar behaviour to instances of the same category, without learning them a priori and, on the other hand, to change its behaviour preferentially based on constraints that produce new categorisations. While this cognitive ability is favourable, one challenge that remains is the (in)ability for generalisation on the robot’s physical exploration, for instance grasping an object in different spatial configurations. In this chapter, it is assumed an “ideal robot”, with enough somatic skills to support the emerging conceptual content of the learning framework. However, one strong attribute of the methodology is that the learning framework has been substantially enhanced to elaborate many types of stimuli from the environment, ergo allowing the interfaced robotic artefact to be optimised independently.

The three experiments evolve at runtime (enabled by the proposed artifice), which emphasises another benefit of language that mitigates the necessity to pre-program the robot’s behaviour. By modelling categorisation, new exemplars can be readily recognised and, therefore, utilised directly to produce similar intentional outcomes as their counterparts, as if they were the same kind of thing. A robot with a pre-determined logic is not able to handle these occurrences appearing speedily in the real world. Or else it would require a significant amount of a priori training. The continuous learning of the robot here is reliant on human interaction (and the evolvement of the environment), which can be mediated by the regular inexpert individual. Understanding language in rich contexts is a long-term goal. In the near sight, however, the embryonic results demonstrated here seem to validate this target rather promisingly. Consider the natural way a parent would teach the child about a new utensil or potable: “What is the can? This is a can. The can is a container”. This affirmation is readily suitable for the child to fill the cup from the can, without watching the parent demonstrate the act. The artifice introduced in this chapter enables the robot to manifest the same behavioural praxis, which requires real-time concept-learning, and the manifestation of the cognitive aspects of categorisation, abstraction, and voluntary control, all facilitated by language.

This chapter investigated a highly influential aspect of language: multilingual contextualisation. To make robots more accessible to a larger audience, they should be able to meet people’s needs despite their singularities. The robot waiter in the second scenario, trained in an English-speaking environment, can serve a potable to an individual who formulates their beverage preference in a different tongue (Italian). Even more appealing is that it needs not to be re-trained to efficiently respond to the same instruction in the new language. This reinforces how linguistic labels, when attached to concepts and becoming part of their representation, have the power to trigger the same mental states as the motor experiences around those concepts. People can easily reuse experiences and skills acquired during their lifetime to unfamiliar encounters. It is desirable that robots do the same, with enough conceptual content and understanding competence. In this chapter, this is demonstrated through meaningful results on how the robot regulates its behaviour autonomously between multilingual conceptual domains. This not only reduces the need to re-train robots in each language, but it also overwhelms the problem that some languages are more disadvantaged than others in terms of learning resources. The ability to transfer earlier-learned experiences as well as to exploit every piece of information available, including eclectic linguistic scenes that are sources rich in semantics and interactions, can be a powerful tool to advance robotics.
Chapter 6

Final Thoughts and Prospect

This dissertation was aimed at modelling high level-cognitive phenomena in robotic artefacts. By high-level cognitive phenomena, it is referred to the human cognitive competencies of learning, categorisation, abstraction, and voluntary control. Humans are remarkable at problem-solving in contexts they do not have direct experience and often involving an act of invention. This is devoted to our conceptual content and development, reflected in the way we represent our knowledge about the world into coherent patterns that allow us to make non-trivial inferences.

The research described in this thesis is inspired by the near absence of high-level cognitive models in robotics. It transcends disciplinary boundaries, from developmental and cognitive psychology, cognitive linguistics, neuroscience, artificial intelligence, and robotics. Most advanced models in the existing literature focus majorly on elementary lower-order cognitive phenomena (perception, manipulation, motor coordination, navigation). While this dissertation preserved the core theory of embodied and situated cognition accounted by these models, it raised an important research question, which they often neglect and what can potentially help transition from low-to-high level cognitive competence of robots. This question governed the research conducted in this dissertation: *where does language fit in the humanlike cognitive development and the emerging vision of the intelligent embodied agent?* Influential insights in cognitive psychology recognise the importance of language in development, in that it complements many forms of neural processing and guides reasoning (*Vygotsky, 1962, 1978*). The rich cross-disciplinary literature suggests that to benefit from language to develop and spur conceptual content, one must demonstrate language proficiency (requisite 1), mastery of grounding it in the world (requisite 2) and the emergence of conceptual content (concrete or abstract, where abstract is a language attribute) to represent knowledge in due ways for further thought and reasoning (requisite 3). Therefore, this thesis aimed to follow the same requisites in stages, in the hope that the modelling of each requisite would generate assumptions for the next, and altogether would complete the emerging picture of the intelligent robot. To address this desideratum, this thesis leveraged a brain-inspired learning framework (*Golosio et al., 2015*) as its skeleton to conduct the intended research, which is compatible with the current knowledge on how our brain processes information.

**Contributions to knowledge**

The study described in *Chapter 3* is conducted in response to requisite 1 and its driven hypotheses. It aimed to contribute to the computational understanding of appropriate dispositions that favour multilingual cognition that is closer to how humans learn, inspired by findings on language control in the bilingual brain (*Crinion et al., 2006*). *Per se*, the study described here was conceived to bring forth potential assumptions on the impact of multilingualism on the learning framework, in the acquisition and discerning competence of multiple languages. The *main contribution* of this chapter is a cognitively sensible multilingual modelling. It offers to the absent literature a model that can liaise in multiple languages simultaneously with no cost of architectural sophistications or augmented demand for computing resources, approximating much closer to human multilingual cognition than powerful large multilingual models (LMs), which learn artificially from engineered
solutions designed for exactly one problem and using myriad corpora. Modelling (humanlike) language development, alone or jointly with other languages, is not a subtle problem, given the languages’ intricacy and the burden of understanding the concurrence of the general mechanisms of information processing in the brain (Dupoux, 2018). The method modelled here demonstrated favourable competence in managing more (four) languages at a time of significant complexity and substantial differences among one another. This could result in huge impact on many cognitive phenomena, such as executive control, speech production and learning, which are currently studied in the human brain (Abutalebi, 2012), and which could potentially be captured in brain-inspired computational models. Indeed, the conducted research sought to study how the central executive could manage interference between languages that are competing parallelly, resulting in sufficient cognitive control across languages and tasks. This thesis marks an archetype for studying and/or explaining language development and the influence of language on cognition, behaviour, and brain structure and functions, according to Hayakawa & Marian, (2019), further propelled and supported by the study in chapter 5, as it relies on cognitively plausible principles over crafted solutions. On the other hand, the conducted study also suggests that this modelling is highly suitable and as effective when deployed in low-cost social robots for human-robot interaction in multiple tongues simultaneously. The modelled robot, able to self-switch its language to match that which is spoken to it can offer a swathe of benefits from both a cognitive and non-cognitive perspective and can help study in-depth the cognitive, neural, and social benefits of multilingualism. Such robots might be more apt in meeting the demands of a larger audience and competently handle endeavours contextually without loss of meaning in translation, a desirable long-term yet not fully attained goal in contemporary AI. Robots can be situated in real environments, where language intricates with multiple stimuli, allowing a more comprehensive study of the cognitive phenomena. Built-in the larger-scope research of the thesis, chapter 3 envisions the imprint of multilingual cognition in high-level processes (e.g., cognitive control), for example how different language stimuli can potentially alter high-level reasoning and decision-making that may result in significant consequences for cognition and the intelligent behaviour of robots.

The research described in Chapter 4 is governed by requisite 2 and its pitched hypotheses. It aimed to radically integrate language in the embodied perspective of cognition to ameliorate learning. An immediate difference of this research with respect to the existing literature is how it views language. Many authors trust that learning journeys from action to language given that word-learning must be grounded in motor modalities. In turn, this chapter sought to demonstrate that language seems to have a similar impact on action-learning as actions have on word-learning. Moreover, some aspects of word learning extend beyond the grounding paradigm, in particular for words that map to abstract concepts that have no direct physical manifestation (e.g., always, never), those that refer to mental states (e.g., think, believe) or used to express intentions, emotions, counterfactuals (if X had not occurred...). While word-learning can undoubtedly emerge from concrete action, there is no substantial evidence that visual/motor context determines the meaning of these abstract words (i.e., contra the action-to-language view). Though this chapter does not commit to investigating such abstractness, it shows that learning, be that of actions, language, or both can journey from concrete concepts directly grounded in action experience to (more) abstract concepts grounded by linguistic experience as well as by their relationship to earlier-learned concepts (continuum). By proposing this learning artifice, two significant consequences emerged, which suggested that:

1. earlier-learned experiences can be used to construct representations for new concepts with the help of language. This snowballs the total learning experience of the model/agent/robot
at runtime, increasing the level of abstractness preferentially, which is rather unattainable in the absence of language.

When modelling a robot’s action-learning, some concrete words can be directly mapped to specific motor instances, like grasp, push, pull. Instead, higher-order (abstract) words can a) have more than one meaning, depending on the desired outcome and context constraints; b) implicate more convoluted motor modalities. By means of a putative continuum, these convoluted modalities can emerge as a combination of primitive (low order) motor instances (datum b), in the way that is dictated by the meaning of the context itself (datum a). For instance, in the manipulation tasks considered in this chapter, to “cut” something was the combined outcome of the action primitives “push” and “pull”. As the abstractness increased, words like “use, make, do” involved more tools and motor modalities that were combined in the current context to obtain the desired outcome. As an example of the meaning of “make”, making tea expected a set of tools and a series of duly combined lower-order operations enacted on those tools (datum b). However, making a toast required a distinct set of tools and interactions (datum a). While “make” mapped onto similar concepts, it also involved different outcomes and modalities to obtain them. The similarity means that the same line of reasoning could be followed for both tasks, which can be, as a minimum, that the task could emerge by dint of combining basic actions from the ground up in a meaningful way that fit the context (outcome, tools, interactions). Therefore, the robot could execute a new task by replicating its earlier-learned experience in a similar context, i.e., abstract from their differences and focus on the procedural similarity. The difference means that the robot needed to solve the task with the new available tools to achieve the new intended outcome, i.e., abstract from the tasks’ procedural similarity and focus on their difference (context-specific constraints). What is worth noting, is that this competence evolved and adapted continuously at runtime learning, extending the robot’s motor map and linguistic experience, without the cost of re-training the entire network de novo for non-anticipated contexts, in contrast to close efforts, e.g., (Stramandinoli et al. 2012).

2. language helps to segment and structure complex observed actions, by describing intricate tasks explicitly in smaller chunks that can be run one at a time.

To illustrate, when following an assembly manual, observing drawings alone might be unattainable to complete the task. Instead, labelling the parts and using verbal descriptions along with sketches boosts the process. Similarly, actions like placing an object on the table in the devised experiments required multiple primitive actions, with no related motor experience and no demonstration of the action provided. Nevertheless, the robot could execute the instruction by simply following the verbal descriptions of the human: to place the object on the table…. Hence, the task could evolve gradually, with creative use of language fitting the workspace (i.e., using it desirably to guide the interaction), instead of the task being a pre-determined end goal of motor modalities. On the other hand, this competence can be opportune when language is the only stimulus available to manifest learning (e.g., describing the action over demonstrating it).

In a nutshell, the main contribution of this chapter can be summarised as follows:

1. It presents the first robotic instantiation of a complex language architecture based on the principle of the Working Memory, to model the symbolic capacity of a cognitive robot to map its internal somatic and mental representations of concepts into their respective terms, starting with those with more evident real-world counterparts (concrete concepts) to those with a weaker explicit or convoluted physical manifestation (abstract notions).

2. It offers an artifice for continuous open-ended behaviour-learning owing to language. It demonstrates how the robot starts with limited knowledge on the task and gradually learns
to explore and manipulate an unanticipated workspace at runtime to achieve the desired
goal, guided by verbal liaison with a human (a learning snowballing).

The research described in Chapter 5 is inspired by requisite 3 and its drawn hypotheses. It is the
culminating study that may respond to the question of whether language can be used as a cognitive
tool in the conceptual emergence of robots. Current advances are unable to grow past their pre-
trained state and poorly using language to propel high-level cognitive competence (categorisation,
abstraction, and voluntary control). The research described in this passage demonstrates significant
contribution in capturing these phenomena and offers to the absent literature a plausible modelling
of conceptual development in cognitive robotics. Concepts allow us to organise the accumulated
experience into coherent patterns and to draw inferences in circumstances in which we lack direct
experience. Thus, analysing a new event is facilitated by assessing the attributes and the constraints
around the concepts that are present in the event. When categorising concepts pragmatically, new
exemplars are readily recognisable and their attributes and or/constraints become apparent, and as
such help to manage further new events. Similarly, the robot in the considered scenarios can learn
from one occasion alone, given that for novel occasions, it can identify new examples of the
concept by means of its categorisation competence, and use them directly without learning those
exemplars (examples of concepts) ahead. For example, if a bottle and a jug share the attribute of
being water containers, learning to use the bottle in a task is sufficient to replicate the same task
with its substituent jug (concept exemplar not introduced when learning the task). It is worth noting
that this ability is solely obtained from verbal descriptions of the concepts (e.g., is a container). It
is undeniable that categorisation can emerge purely from vision, however, the number of categories
is much less than those in the presence of language. Language grants social cues that are generally
unobservable (e.g., our emotions) and can generate ever-increasing forms of categorisation. The
linguistic labels that we associate with our conceptual representations can be used to communicate
those representations to other people in social contexts. For instance, by means of observation, a
dog and a fish belong to distinct classes. Yet, if we say, “I have a dog” and “I have a fish”, the dog
and the fish become part of the same category of pet animals, which can only emerge in response
to language over perception. Similarly, a bottle and a jug may look physically different, but they
can both be used as containers. Yet, sharing this attribute does not always place them in the same
category. For example, though we can observe that a water bottle and a wine bottle look similar
but differ in colour, it is the linguistic stimuli to determine whether they belong to the same
category (e.g., when sharing the attribute of containers) or not (e.g., depending on the context: “to
make tea you need water” – water bottle, “Mary likes wine” – wine bottle), which shapes the
robot’s voluntary control. Language does not only help build abstract words (continuum,
snowballing) but as per Clark, (2006) it produces knowledge structures that themselves represent
instances of perception, manipulation, and (further) thought.

One highlight of this study is the multilingual cognitive competence demonstrated in scenario 2 in
the conceptual content of the robot and how it influences problem-solving. This method breaks
through the traditional machine translation, directly exploiting conceptual content to arrange cross-
lingual retrieval. By making linguistic labels part of the concept representation, concepts and the
experiences that surround them can be readily accessed in new situations, for example, when a
concept is related to an earlier-learned motor experience, its label triggers the inner representation
of the concept that allows reproducing the mental states of the motor experience. When this label
is of a different language, the motor experience can be replicated without learning it explicitly in
that language. Thus, by contextualising events in poly languages, this study aimed to investigate
whether they (events) can be understood by analysing the constraints and concepts that surround
them, without re-learning the concepts or grounding them with new semantics. Multiple languages
involve a rich semantic memory for concepts and further social cues that can influence reasoning and the decision-making process. The brief idea behind the research described here, is that a robot with solid *experiential knowledge* in one language (say English) and little but sufficient *linguistic knowledge* in another language (say Italian), will be able to exploit its earlier-learned praxes (in English) to self-regulate the accomplishment of goals in the unfamiliar environment (Italian), in which it does not have direct behavioural experience. This results in:

a) linguistic terms (in Italian) draw attention to the appropriate concepts that surround an event, allowing the event to develop with little learning by exploiting the accumulated experiences with those concepts (in English). Using solely linguistic cues, the robot does not need to observe or be guided through the task *de novo* to produce the desired behaviour in the new semantic environment. Instead, language alone activates the same (mental) states as the motor experience (pertaining to the concept) and allows to transfer that experience onto the new semantic context.

b) linguistic cues can bootstrap the creation of novel representations in the new language and thus grasping the intricacies of that language (Italian) through context and experience (English), e.g., learning what “servire” (*to serve*) a drink means in terms of real-world manifestation.

*How else can we benefit from two language knowledge sources?* From the perspective of concept understanding as defined by the learning sciences, a learner must be able to identify the concept and associate its criterial attributes accordingly under unfamiliar circumstances. It is unattainable to program robots with all requisite knowledge necessary for problem-solving and, nevertheless, this cannot anticipate the novel endeavours that occur in the dynamic real-world environment. Thus, our best effort is to equip robots with “innate” skills to exploit every piece of information from all available sources that exist or are presented during their lifetime. That includes diverse linguistic environments that are rich in semantics and interactions. To utterly understand concepts, these models must identify them suitably across the different semantic predicates to retrieve their meaning/representation autonomously and readily. Moreover, two (multiple) human languages comprise different knowledge sources, which can help disambiguate incomplete workspaces. For instance, when information is not available in one language to solve the task at hand, it can exist in another language, either in verbal (e.g., the preferred potables uttered in Italian) or experiential form (e.g., the behaviour trained in English, but not in Italian). The ability to merge the two could propel robot learning even under uncertainty and help bootstrap experiences in languages that have poorer resources of learning corpora.

**Contributions to the learning framework**

The research described in this dissertation offers not only significant contributions to knowledge but to the learning framework as well. Such refinements are central to accommodate the type of investigation and modelling performed here and come in the following forms.

1. One component of our working memory model that has not been exhaustively discussed in this thesis is the *episodic buffer* (*Baddeley & Hitch, 1974*), which corresponds to the *focus of attention* in the model of *Cowan, (1995)*. This structure is responsible for binding information from several domains (auditory, visual, spatial) in units of chronological order. In the learning framework, the episodic buffer is equivalent to the word group (up to four words as per the proposal of Cowan) that is readily accessible for memory retrieval and the chronology is maintained between the word group and the long-term memory. In the original framework proceeding this thesis, there was only a *verbal* episodic buffer. The
framework also lacks a visuospatial structure that is responsible for handling non-verbal stimuli. Inspired by the view of Cowan, (1995), in which there is no distinction between the phonological loop and the visuospatial sketchpad, the research described here overcomes the lack of the visuospatial structure in the framework, by introducing a new conventional way of representing non-verbal stimuli using lexicalised symbols. These symbols are human interpretable, making it easier to supervise the output generated by the framework and validate if it is the desired outcome (note that the framework learns preferentially without human supervision, only the input and output are loosely monitored). Therefore, the proposed novel episodic buffer of the enhanced framework can handle multidomain multimodal information, which although does not introduce any changes to the global organisation of the framework, radically transforms its procedural learning mechanisms and memory retrieval.

2. The artifice of representing non-linguistic stimuli in a convention of lexicalised symbols (_word) proposed in this thesis has a significant and strategical role in conceptual content. This convention (_word) approximates manifold stimuli that can activate a conceptual representation in the human brain (visual, somaesthetic, auditory, olfactory and taste) in a simplified computational model that does not account for all biological aspects from the neuroscience perspective (the actual manifestation and representation of these stimuli). For example, the concept of a dog can be activated by looking at a dog, hearing a bark or the word dog itself. Similarly, the concept of sweet can be activated by tasting something sweet or saying, “this food tastes sweet”. Moreover, each of these stimuli activates the attributes of the concept and links indirectly to the other stimuli (looking at a dog recalls the bark and vice versa). Therefore, the conceptual content of the robot can be organised in structures of several domains, each of them representing and triggering instances of perception (from multiple senses) or manipulation and account for a range of inputs that influence the behaviour of a robot in the real workspace.

3. In the original framework, words were represented as orthogonal vectors. By introducing the symbolic representation (_word) as a pivot concept between poly language semantics, the orthogonality is broken. The same symbol _word will link to two different labels, e.g., _dog dog, _dog cane (where cane is the Italian word for dog). The two different linguistic labels will trigger the same internal representation of the concept (or mental states), and in turn, the concept representation will activate two linguistic labels competing in parallel. This has two important implications:

- In the original framework, word synonyms or meaning similarities were not identified as words/phrases were orthogonal among one another. This is overwhelmed by the proposed symbolic representation, e.g., the joint representation _dog brings adjacent the synonymous terms of dog and cane (Italian for dog), which is demonstrated by the fact that the robot can transition from one language to the other (scenario 2, chapter 5).

- Breaking the word/phrase orthogonality allows different words/phrases to trigger the same mental action sequences (mental states) that relate to accumulated perceptual and motor experiences. Therefore, the robot can activate those mental states to reproduce the experiential knowledge without training a new mental action sequence for that behaviour with semantics from a different language, which would be the case should orthogonality be maintained. This type of organisation of the conceptual content is the source of the high-level competencies demonstrated in scenario 2 of chapter 5.
This significant assumption (of orthogonality) can be further investigated analytically, via careful examination of the mental action sequences that are generated in the absence and presence of the multilingual symbolic representation proposal.

**Quo Vadis**

Understanding language, disambiguating unanticipated workspaces, continuous adaptive learning from real-world interactions and high-level cognitive modelling are long-term goals in AI/robotics. The computational modelling offered in this dissertation represents a rather promising embryo for further cognitive robotics research to achieve these goals. This thesis aims to inspire more work that integrates insights from theoretical and empirical developmental/cognitive psychology with robotics to recognise language as means to empower embodied and situated cognition, currently studied apart. To shape the conceptual development of robots, it is important to first understand fully that of humans, for example how knowledge is acquired through experience and that acquired through language change and ally in the course of development from early to later stages of learning. The complexity of learning and input considered in the experiments devised here were merely developmental intended for neural processing capacity of pre-school children. It is rather challenging to explore how this can intricate further to obtain more abstract reasoning. In turn, the level of abstractness considered here is moderate, where an abstract word is one that can be segmented in more action primitives. This thesis does not consider the emergence of purely abstract words, e.g., *peace, feel, think*. However, it opens a path to try to explore abstractness at various increasing levels, in a continuous toolchain (*learning snowballing*).

One fundamental limitation that impedes the increase of task complexity rests with the somatic properties and action generalisation skills of social robots. While the cognition of the robot improves, it cannot be fully supported by its motor capabilities, for example in the creation of convoluted modalities for higher-order actions expressed by abstract words. Not only is the motor competence limited in humanoid or low-cost platforms (poor manipulation skills) and energy inefficient (overheating), but the ability of action generalisation is currently attributed only to powerful industrial robots. Hence, the method of learning snowballing introduced in this thesis could converge much faster if coupled with robots that support sophisticated manipulation in the real world (over simulations). Another appealing path that can be explored in a similar fashion is robot navigation in real scenes, by means of integrating language with the spatial representation of the robot’s workspace to study how the robot can explore the scene by following verbal instructions (at a different level of abstractness) of chronological sequence until the final target is reached.

A major aspect of the novel methodology proposed in this thesis on the symbolic capacity to map language with other stimuli from the physical environment is the workaround of the absent visuo-spatial sketchpad in the original learning framework. Instead of modifying the architecture to include this as a separate structure, this thesis proposed a symbolic representation of the external stimuli, introducing instead substantial changes to the episodic buffer (that binds the phonological loop with the visuospatial sketchpad) and the learning and memory retrieval mechanisms. While this proposal finds support in some influential theoretical models of the working memory (*Cowan, 1995*), a question that may deserve further investigation could be *is it intrinsic to devise a separate visuo-spatial sketchpad structure and in what way would it affect learning contra the methodology offered here?* This question could inspire future research on cognitive architectures compliant with the principles of the working memory (some surveyed in 2.2.2.3) or to modify the current learning framework accordingly to capture these pressing aspects. A putative path to take could be to apply
the artifice of symbolic representation offered here in other similar complex frameworks and verify its potential, on its own and in comparison with other proposals, ideally those that include a visuospatial sketchpad as a separate component to represent the physical environment of the robot.

A prosperous research path this thesis lays a stone of is exploring the consequences of multilingual experience in cognitive architectures. That would require to study in-depth how inputs of different languages affect the levels of processing. Multilingualism is becoming the norm over the exception inspiring giants like Facebook/Meta to invest (in 2022) in a simultaneous language translator that would allow people to communicate with one other while each speaking their own tongue. This is a vision that this dissertation had early in its genesis (2019) when modelling the bilingual cognitive robot described in passage 3.5. Multilingualism will continue to nourish novel ideas on potential applications with an immediate imprint in our society and the emergent picture of cooperative AI/Robotics. From a cognitive research angle, further modelling of multilingual cognition in brain-inspired architectures could reveal important insights on how the architecture accommodates different languages and what procedural and/or structural changes (if any) might occur from the parallel handling of more than one language. For example, an in-depth investigation of the neural activations during simultaneous multilingualism in the learning framework considered here could help understand how language conflict is overwhelmed and if it has any influence on its cognitive functions, potentially comparing with other more realistic biological models.
Appendix A

The learning framework

I. Description of the cognitive architecture

This section describes more in-depth the learning framework leveraged in this thesis introduced in section 2.3.4.1. The global organisation of the model is illustrated in Supplementary Figure 1. It comprises four major neural structures: a short-term memory (STM), a long-term memory (LTM), a central executive (CE) and a reward structure (RW). The long-term stores explicit knowledge (phrase memorisation structure) and is responsible for phrase retrieval. In the original version of the framework, each memorised phrase in the LTM corresponds to linguistic stimuli only, whereas, in the contributed framework in this dissertation, phrase (symbolic) can be either linguistic, visual, somatosensory information or a combination (multimodal, multidomain). Phrase retrieval from the LTM is supported by the focus of attention (STM) that holds readily accessible information (up to four words/symbols) used as a cue to pivot to relevant information in memory. A notable structure of the short-term memory (STM) is the phonological loop. This maintains the working phrase that is currently processed by the model through the mental actions. The goal stack (typical in cognitive architectures) contributes to decision-making indirectly, by holding goal chunks (an action kept as a future goal), when that action cannot be executed immediately. The information contained in these two structures (phonological loop, goal stack) is coordinated by the comparison structure. In other words, this component recognises similarities among the phrases in the phonological loop, the goal stack, and the focus of attention (cue structure), to retrieve the information that best applies to the desired decision-making.

![Supplementary Figure 1](image-url)
diagram. The arrows that join the rectangles represent directional connections among neurons of different subnetworks and buffers. During the acquisition phase, the input words are mapped onto neurons of an input-word buffer and are combined into an input phrase through proper indexing. In the association phase, the input phrase is copied to a working-phrase buffer in the phonological loop and is mapped to a neuron of the memorised-phrase buffer. The phrase is stored in the long-term memory (LTM) by changing the weights of the connections from that neuron to the working-phrase buffer, through a discrete version of the Hebbian learning rule. During the exploration and exploitation phases, word groups are extracted from the working phrase buffer and used by the focus of attention as cues to retrieve memorised phrases from the LTM. When the working phrase represents an action in the mental action sequence that cannot be performed immediately, it is copied to the goal stack, where it is stored until the action is carried out. The comparison structure is a subnetwork that recognises the presence of similar (equal) words in the working phrase, the phonological loop, and the goal stack. The elementary operations (mental actions) on word groups, phrase buffers and other SSMs (sparse signals maps) are triggered by the action neurons. The action neurons control the state of the gatekeeper neurons, which, in turn, manage the flow of signals between the subnetworks by acting as gates that allow or inhibit the information to pass. When the exploration leads to the desired output, the human interlocutor can trigger a reward. In the reward phase, the system retrieves the action sequence that led to the reward and memorises the association between the state of the system in each step of the sequence and the corresponding action. The state-action association is memorised by changing the weights of the input and output connections of the state-action association subnetwork through the k-winner-takes-all rule.

A versatile component of the framework is the central executive (CE). The CE is a neural structure that controls all statistical decision-dependent processes inside the framework. It includes a state-action association (SAA) network, the action neurons, and the gatekeeper neurons. They are based on the same neurological model but differ in how they connect to other sub-networks. The action neurons are used to perform basic operations on words, word groups, or phrases, such as acquiring words from input phrases, memorising phrases, extracting words from a working phrase, retrieving memorised phrases using cue word groups. Action neurons are connected to gatekeeper neurons using pre-determined fixed-weight connections. When action neurons are activated in response to mental actions, they trigger the connected gatekeeper neurons. Given that the neural connections are fixed, specific action neurons will habitually fire the same gatekeeper neurons connected to them. The gatekeeper neurons are fully connected to one or more sub-networks. When these neurons are fired, they allow the flow of signal through the sub-networks and inhibit it, otherwise. Consequently, by controlling which gatekeeper neurons are activated, the action neurons manage the flow of information among the slave systems (neural structures). In turn, this ensures that specific operations executed by action neurons will always activate the same state of the network. The gating mechanism is controlled by the state-action association (SAA) neural network. This associates the mental operations executed at neural level with the model’s internal states (the pool of the neurons that fired in response to these mental operations). When this association has led to a valid (desired) output, the reward structure sends a reward signal to the SAA neural network to memorise the association permanently. If in future endeavours the model will receive an input signal that is similar to a trained input, the SAA will retrieve the memorised association and the model will execute the same mental operations that lead to a valid output for that type of input.

A fundamental trait of the procedural learning mechanisms is that the learnable connections, i.e., the connections affected by the reward structure, are connected to the action neurons rather than being directly connected to output words or phrases. Therefore, the model learns to build the output preferentially, following autonomously a neural level procedural sequence of relevant operations on phrases, rather than learning those words, phrases, or their word/phrase combinations to solve the task. This enables it to handle several tasks instead of specialising in ad-hoc input and, unlike traditional state-of-the-art deep-learning models, to learn complicated rules from a few examples, akin to how humans learn naturally and with reasonable accuracy from only a few occasions. The
source code of the learning framework software, a comprehensive User Guide to use the model and original datasets used for validation are available at https://github.com/golosio/annabell/wiki.

II. The learning principles and mechanisms

The ANNABELL neural network architecture design is based on the concept of the sparse-signal map (SSM), which is an artificial neural network with only a small fraction of neurons being active at a given time. The architecture consists of several SSM subnetworks, which, in turn, are composed of interconnected artificial neurons. The subnetworks are generally connected through variable-weight (learnable) links. The inhibitory competition among neurons is modelled using the k-winner-take-all rule, i.e., the k neurons with the highest activation state are switched on, while all the remaining neurons stay off, in combination with a discrete version of the Hebbian learning (DHL) rule. The k-winner-take-all rule provides a computationally effective approximation of the activation dynamics produced by inhibitory interneurons. Moreover, this rule is biologically justified, in contrast to the standard backpropagation learning algorithm, which is however used by many biologically inspired neural models of language as it grants error minimisation.

The Hebbian theory describes a theoretical learning mechanism based on synaptic plasticity in biological neural networks. This theory underlines the principle of neurons that fire together, wire together, i.e., the strength of their synaptic junction is reinforced when the neurons are repeatedly active, or their outputs are strongly correlated (Hebb, 2005). In the ANNABELL model, this means that if the pre and postsynaptic neurons at the opposite ends of a connection are in the same output state (above or below threshold) the link weight is saturated to its maximum value. Otherwise, it is saturated to its minimum value. Although the discrete-Hebbian-learning (DHL) rule is an extreme simplification compared to other models focused on biological realism, ANNABELL is based on the same learning principle that is responsible for synaptic plasticity in biological neural networks. A more realistic implementation of the Hebbian learning rule, with small updates of the connection weights, would require significantly large computational resources for training and validating the framework on a large corpus, which would largely affect the real-time interaction with the end-users.

III. The mental actions and the neural gating mechanism

The synaptic gating theory postulates that neurological gates in the cortex and other brain areas allow or inhibit nervous stimuli in the brain (Gisiger & Boukadoum, 2011). Theoretical and empirical findings link the neural gating mechanism with the human capability to filter out irrelevant information and retrieve relevant information from the working memory (McNab & Klingberg, 2008). In the ANNABELL framework, the neural gating is simulated by the gatekeeper neurons, connected to the action neurons via fixed connections. The neural gating is a prominent mechanism of the framework, which grants its adaptive learning and generalisation competence and is implemented as follows:

The gatekeeper neurons are generally fully connected to a set of bistable neurons (Supplementary Figure 2). When the gating signal is “off”, the state of gated bistable neurons is “down”, hence they are not triggered by the input signal. When the model performs a specific elementary operation, an action neuron is fired, which sets the gating signal “on”. The open gate allows the gated bistable neurons to change in the “up” state and transmit the input signal to the second set of neurons. Given that the connections between the action and the gatekeeper neurons have fixed predetermined weights, each action neuron will only correspond to the specific mental operation that triggered its activation.
The mental action sequence

In classical tasks used to study working memory capacity (Oberauer, 2002), a subject is asked to memorise a short, varied sequence of digits and perform elementary arithmetic operations on the sequence or on a subset. Let there be the following task: “Add one to each of the following digits: 7 2 4 3 9”. To simplify the discussion, it is assumed that the model has some initial knowledge on simple arithmetic additions, acquired from past experiences. Affirmations like “A plus B equals C” are pre-memorised in the LTM to keep the cognitive load small.

To solve the task, the model must undergo the steps below:

1. Transfer the phrase “Add one” to the phonological store;
2. Transfer the phrase to a goal-stack store;
3. Transfer the given sequence of digits 7 2 4 3 9 to the phonological store;
4. Transfer the first digit (7) to the focus of attention;
5. Retrieve relevant information stored in the LTM, using the digit as a cue and transferring it to the phonological store. Relevant information in this scenario can be a sentence in the LTM that is appropriate for the requested arithmetic operation, e.g. “one plus seven equals eight”;
6. Transfer the result (eight) to the focus of attention and use it in sentence production (verbal or speech);
7. Transfer the initial sequence of digits 7 2 4 3 9 to the phonological store;
8. Transfer the second digit (2) to the focus of attention;
9. Execute the same mental actions from retrieval to sentence production and repeat iteratively until the last digit of the sequence is processed;

Mental operations a and b comply with several studies (Bryck & Mayr, 2005; Vandierendonck, 2012) which suggest that the task goal should be stored in the working memory in some directly accessible form. A minimal system that can perform this sequence should include (at least) the following components: a phonological store; a focus of attention; a retrieval structure from LTM; a goal store and a supervising system that controls the flow of information among the other
components, i.e., a central executive. It is necessary to design the working memory as a neural architecture over a symbolic system, not only given the naïve and obvious consideration that the human brain is a neural architecture but because the decision processes operated by the central executive, are statistical decision processes, as opposed to rule-based processes (managed by the central supervisory component). When probed to execute a similar task with different digits "add two to each of the following digits: 8 7 2 1 5", the model must retrieve the action sequence (a-b, 1-6). When performing an addition with the first digit 8, the model might encounter any of the following phrases in the long-term memory:

"eight plus one equals nine" (phrase 1)
"eight plus two equals ten" (phrase 2)

Should the retrieval decision among phrases 1 and 2 be based solely on similarity score between each of the phrases with the phrase retrieved when learning the task ("seven plus one equals eight"), the system would pick on phrase 1 as the most appropriate and generate the wrong answer “nine” instead of “ten”. However, given that the task is transferred to the goal store (operations a-b) until the end of the decision process, the comparison structure of the architecture will recognise that phrase 2 and the goal phrase “add two” contain the word “two”. Since the comparison structure is also a neural network, upon identifying the similarity, the neurons that compare the two words will be activated. In the model, the connections from the comparison structure to the central executive have greater weights than those from the phonological store to the central executive. When the neurons of the comparison structure are activated, in the considered example the model will select phrase 2 over 1 and give the correct answer “ten”. This emphasises why the retrieval process should be modelled using a neural architecture, or more generally why the retrieval process should be described as a statistical process.
Appendix B

Examples of interaction with the framework

I. The Interface

The framework and the human interlocutor communicate through a user interface. The interface:

1. Converts words and phrases into ASCII representations and submits them in sequential order (one by one) to the system.
2. Converts ASCII representations generated by the system into words.
3. Sends reward signals to the system when a reward command is prompted by the human interlocutor.

Training the framework consists of two parts:

1. Using the interface (or a text file, which is submitted to the system through the interface) the user types several declarative phrases - verbal information that the system should memorise in its long-term memory (referred to as Declarative Knowledge, see Chapter 3).

2. The human interlocutor types an input phrase that the system will learn to answer. Learning how to answer the input phrase consists in executing a number of sequential commands, which prompt the system to perform elementary operations at sentence level for that (specific) input phrase to build a valid output. The series of elementary operations executed by the system (triggered by the user commands) generate a mental action sequence. When the system has built a valid answer for that type of input (i.e., the output desired by the user), the system rewards the system by sending a reward signal through the interface. This prompts the system to memorise the mental action sequence that is generated and to permanently memorise the association between the action sequence and the internal state of the system (activated neurons and all updated learnable connections during training). This is referred to as Procedural Knowledge, Chapter 3.

Note that, training the system to produce a valid answer depends on the type of the requested input phrase. That means each time the human interlocutor chooses to train a new input phrase, the series of elementary operations (i.e., interface commands sent to the model) will change depending on the desired output. These commands are appropriately selected by the user to reach the desired goal. For example, the answer to the input phrase “what is your sister’s name” (output: her name is Susan) will be different from that of the input phrase “do you have a brother” (output: yes, I do) and, hence, training these inputs requires the use of a different sequential series of commands.

Testing the framework:

1. Using the current state of the trained system, the user can type through the interface other input phrases similar to those trained, followed by the respective command that prompts the system to start an exploitation phase on the input (see The Commands).
II. The Commands

The (main) commands used to train the system are listed below. Commands start with a dot and can be abbreviated (as illustrated inside the parentheses).

**annabell**: launches the main program

**phrase(.ph)** *phrase*: prompts the system to retrieve a phrase from the association process

**word_group(.wg)** *word group*: prompts the system to extract the given word group argument from the working phrase

**push_goal (.pg)**: inserts the word group at the top of the goal stack structure of the architecture. A word is inserted into the goal stack when an exploration cannot be performed immediately.

**sentence_out (.snto)**: sends to the output the part of the sentence that comes after the current word group and all the subsequent phrases of the same context (a group of phrases that make a paragraph). In turn, it rewards the system.

**partial_reward (.prw)**: sends a partial reward to the system and the word group to the output

**reward(.rw)**: sends a conclusive reward to the system – the system sends the last word group (before the reward) to the output and permanently memorises the state-action sequence for the current input phrase that it is learning.

**exploit(.x)** *phrase*: prompts the system to start an exploitation phase (during the test). The system is expected to generate a valid output that appropriately responds to the input phrase, which is being tested.

**file(.f)** *filename*: this is an alternative way of interacting with the system – instead of typing the commands one by one in the interface, they can be written in a text file, which is called in the interface. The system loads the contents of the file line by line and executes the specified commands accordingly.

**save** *filename*: the current trained state of the system (weights of connection links) is saved in a file and can be loaded for future testing of the system without retraining.

**quit(.q)**: is used to exit the program. If not saved, all connection weights of the system are erased and the next time the system is launched, training will begin from a clean slate.

III. Training Example 1: Simple Phrases (Chapter 3)

In this example, the human interlocutor is guiding the system to reply to the question “how old is your sister”. The desired output is “she is seven years old”. The user needs to execute a series of commands through the interface that will lead to this output. This example refers to Algorithm 1, Chapter 3.

Stage 1: Training the declarative knowledge

Initially, the human user must feed into the system the necessary information that will be used to respond to the input phrase. Such information might include the following declarative phrases:

*the personal pronoun for a girl is she* | *the personal pronoun for a boy is he* | *the personal pronoun for a woman is she*

*Susan is your sister* | *Susan is a girl* | *Susan is seven years old* | *Susan goes to school*
Submitted Thesis: *Human Language as a tool for Conceptual Development in Cognitive Robotics*

Mike is a boy | Mike goes to kindergarten | Mike is your cousin | Mike is three years old  
Alice is your mother | Alice is thirty years old | Alice is a journalist | Alice is a woman  
… (other similar phrases that include different people’s names, professions, ages, …)

Note that the order in which the phrases are submitted to the system is not important. The system may have memorised other unrelated phrases along with the ones above. Here, these sample phrases are given in relation to the input phrase that is being trained.

**Stage 2: Training the procedural knowledge**

The human interlocutor will ask the question (through the interface): “*how old is your sister*”.

To train what answer will be produced by the system, the user must suggest (using appropriate commands) the target words and phrases that will be manipulated to build that answer. Hence, the user will guide the system using the following commands:

`.wg your sister`

This prompts the system to start an exploration phase that extracts the word group *your sister* from the working phrase buffer.

`.ph Susan is your sister`

Starts a further exploration phase to retrieve the sentence *Susan is your sister* from the long-term memory through the association mechanism.

The user intends that the system uses personal pronouns to refer to people, instead of their names. Henceforth, the user will execute the following command to obtain the desired word group extractions and associations.

`.wg Susan`

Exploration to extract the word *Susan* from the working phrase (*Susan is your sister*)

`.ph Susan is a girl`

Exploration to retrieve the phrase *Susan is a girl* from the long-term memory

`.wg girl`

Exploration to extract the word *girl* from the working phrase buffer

`.ph the personal pronoun for a girl is she`

Exploration to retrieve this phrase from the long-term memory

`.wg she`

Exploration to extract the word *she* from the working phrase buffer

`.prw`

Send the word to the output and (partially) reward the system

The word group *she* obtained from the exploration phase is the first part of the output that the system will generate. The partial reward command (*.prw*) prompts the system to send the word group at the output buffer but announces to it that the output phrase is not complete. The final output is only created when a full reward (*.rw*) is given.

To complete the answer, the teacher must use the following commands:

`.wg Susan`

Exploration to extract the word group *Susan*
.ph Susan is seven years old
Exploration to retrieve this phrase from the long-term memory

.wg is seven years old
Exploration to extract the word group from the working phrase buffer

.rw
Send the word group to the output and (fully) reward the system

The final output is now created. To validate, the teacher tests the input phrase as follows:

how old is your sister

.x
The (.x) command starts an exploitation operation mode, which attempts to build an answer. Should the system respond, “she is seven years old”, the training has been successful.

*TEA: how old is your sister
*SYS: she is seven years old

Where, SYS = ANNABELL system; - TEA, Teacher.

The user can test the system with similar input phrases. For example:

how old is your mother
how old is your cousin

The system should answer “she is thirty years old” and “he is three years old”, respectively. The system can use both pronouns (he/she) even if trained only for the feminine gender, by means of extracting the word groups that cue the relevant declarative phrases in the long-term memory, associated with those groups.

As noted previously, the sequences of training commands selected by the user (along with the word group/phrases extractions and associations) are specific to the trained input phrase. For example, the type and sequence of commands to train a different input could be:

**Stage 1: Training the declarative knowledge**

Relevant declarative phrases

Susan ’s favourite game is Pictionary | Susan ’s favourite cartoon is Candy Candy | Dad ’s favourite music is classical music | Mum ’s favourite singer is Bryan Adams | Oliver ’s favourite game is football | Vivian ’s favourite cartoon is Heidi | Letizia ’s favourite cartoon is Peppa Pig | ...

**Stage 2: Training the procedural knowledge**

what is Susan ’s\(^5\) favourite cartoon

.wg Susan

.ph Susan is a girl

.ph the possessive pronoun for a girl is her

.wg her

.prw

.pg Susan

---

\(^5\) By convention, words with suffixes or apostrophes must be split in the form: base – suffix, base ’s (e.g., writing must be submitted as write-ing; Susan’s must be submitted as Susan ’s).
The output of the exploitation phase would be:

*TEA: what is Susan’s favourite cartoon
*SYS: her favourite cartoon is Candy Candy

In the generalisation test, the human interlocutor can ask similar questions involving different people (and their specific preferences). For example:

*TEA: what is Dad’s favourite music
*SYS: his favourite music is classical music

*TEA: what is Oliver’s favourite game
*SYS: his favourite game is football

The system can answer about any gender and any preference, using sentences of variable lengths (the .sentence_out command allows producing the remaining part of the sentence after the word group).

The property that enables the system (learning framework) to generalise and abstract in this way is devoted to its learning mechanisms. The input of the state-action association structure (SAA SSM – sparse signal map) includes equal-words vectors. When the system is trained to extract the word group favourite cartoon and insert the word group Susan in the goal stack, it will recognise that the phrase Susan’s favourite cartoon is Candy Candy is more appropriate than any of the phrases Susan’s favourite game is Pictionary or Vivian’s favourite cartoon is Heidi, considering that the word group (favourite cartoon) in the sentence Susan’s favourite cartoon is Candy Candy is equal to the extracted word group favourite cartoon and the last word group in the input sentence (what is Susan’s favourite cartoon). At the same time, it will recognise that the word group Susan pushed in the goal stack is equal to the word group Susan in both phrases Susan’s favourite cartoon is Candy Candy and what is Susan’s favourite cartoon. (Note that equal words are identified by their vector representation. Each time that word is submitted into the system, it will have the exact same ASCII representation). By learning this association from the training example, the system can generalise following the same line of reasoning for other inputs of the same type and produce valid answers.

IV. Training Example 2: Snowballing Toolchain (Chapter 4)

This example illustrates a sample construction of the snowballing toolchain.

Stage 1: Training the declarative knowledge

The human interlocutor types through the interface (or loads a text file) a series of declarative phrases (factual knowledge) that are needed to complete the task. Such phrases may contain verbal instructions (e.g., the bottle has water or to make tea you need a mug, teabag, and water) and can be considered as an experience that the learning framework has stored in the long-term memory.

In contrast with the previous learning example (Training Example 1), the information stored in the long-term memory is multimodal, i.e., it contains pseudo-sensory vision and motor data along with
linguistic data. The pseudo-sensory information is represented using a symbolic representation
convention (see Chapter 4). Henceforth, the human interlocutor must populate the long-term
memory of the framework with the sensory-semantic representations (<form form> associations).
These are generated following the proposed artifice of Object Learning and Action Learning
explained in Chapter 4, which allows the framework to self-populate its long-term memory from
the runtime interaction with the human in the workspace. Here, we assume that this knowledge is
already stored in memory as follows:

 grapple grasp
 lift lift
 pick-up pick-up
 drop drop
 throw throw
 cut cut
 mug mug
 bottle bottle
 teabag teabag
 cup cup
 orange orange
 knife knife
 bread bread
 ham ham
 cheese cheese

Stage 1: Training the procedural knowledge

The human will guide the (system) learning framework to convert human verbal instructions of
sequential primitive actions into symbolic representations of actions. The instructions begin with
an indicator (step 1, 2, 3) to maintain sequence order. An example of interface commands used by
the human is illustrated below (note that the commands used here follow a simplified convention,
see Simplification of the training procedure at https://github.com/golosio/annabell/wiki/User-
guide).

step 1 grasp the mug
The requested input phrase

 .po* step 1

Extract and send the word group step 1 to the output and reward the system partially

 .po* /mug/_mug mug/_mug/

Extract the word mug, retrieve the _mug mug association from the long-term memory, extract the
word group _mug and send this word to the output, partially rewarding the system

 .o /grasp/_grasp grasp/grasp/

Extract the word grasp, retrieve the _grasp grasp association from the long-term memory, extract
the word group _grasp and send this word to the output along with a conclusive reward signal.

The teacher validates this learning by typing the following exploitation command.

step 1 grasp the mug
 .xm
The system will generate the following output and will memorise it permanently in memory (.xm = exploit memorise command stores the output permanently in memory after exploitation).

*TEA: step 1 grasp the mug
*SYS: step 1 _mug _grasp

The sequence step 1 _mug _grasp is sent to the robot, which interprets the elements one by one. The word group _mug (symbolic representation of the object) will trigger the object recognition module to track the object and the word group _grasp (symbolic representation of the action) will call the planned action trajectory. The robot will grasp the mug.

As explained earlier, this example is sufficient to generalise to similar instructions with other objects and actions. Hence, the teacher can interrupt the learning and start a testing exploitation phase with the (.xm) command to memorise the output in the framework’s long-term memory. To introduce higher-order verbs, the human interlocutor provides a description of the action followed by the lower-order actions that compose it. At the same time, the human prompts the system to execute the lower-order actions, by starting the exploitation phase (.xm), as follows:

(test, primitive actions)

to take the mug
step 1 grasp the mug
  .xm
step 2 pick-up the mug
  .xm
to place the mug
step 3 drop the mug
  .xm
to take the bottle
step 1 grasp the bottle
  .xm
step 2 lift the bottle
  .xm
to add the water
step 3 pour the water
  .xm
to take the teabag
step 1 grab the teabag
  .xm
step 2 lift the teabag
  .xm
to add the teabag
step 3 throw the teabag
  .xm

The outcome of this test in the human-system interface will appear as follows: (Note that this is not a learning phase; however, while executing the instructions, the system memorises the outcome in long-term memory simultaneously, hence expanding its repertoire of motor exploration. These experiences will be used impromptu in further contingencies).

*TEA: to take the mug
*TEA: step 1 grasp the mug
*SYS: step 1 _mug _grasp
Appendix B

TEA*: step 2 pick-up the mug
*SYS: step 2 _mug _pickup
*TEA: to place the mug
*SYS: step 3 _mug _drop
*TEA: to take the bottle
*SYS: step 1 grasp the bottle
*SYS: step 1 _bottle _grasp
*SYS: step 2 lift the bottle
*SYS: step 2 _bottle _lift
*TEA: to add the water
*SYS: step 3 pour the water
*SYS: step 3 _water _pour
*TEA: to take the teabag
*SYS: step 1 grab the teabag
*SYS: step 1 _teabag _grasp
*SYS: step 2 lift the teabag
*SYS: step 2 _teabag _lift
*TEA: to add the teabag
*TEA: step 3 throw the teabag
*SYS: step 3 _teabag _throw

Each time the system generates an output, this output is interpreted by the robot that executes the action on the referent object, before moving along the toolchain.

After having dictated the action primitive, the human interlocutor can introduce more higher-order instructions. A learning example of a higher-order instruction is illustrated below.

(training, 1st higher order)

Nao take the mug
.wg /mug/_mug mug/_mug/
.po _mug
.rip
.pg /mug & take/to take the mug//6
.wg /mug/_mug mug/_mug/
.ph /_mug/step 1 _mug _grasp/_grasp/7
.prw
.rw
.pg /mug & take/to take the mug//
.wg /mug/_mug mug/_mug/
.ph /_mug/step 2 _mug _pick-up/_pick-up/
.rw

The human can test the framework on the other objects/actions, whilst memorising the outcome of the exploitation phase, as follows:

6 Note that the human inserts the word groups take & mug in the goal stack. The model will compare the words with the ones in the high-level description provided by the human to take the mug to retrieve the lower-order primitive actions associated with that verb.

7 The model re-uses its earlier-learned experience from the earlier exploitation phase and recalls it in the current task. If the model does not memorise this phrase permanently during the exploitation (test), the phrase cannot be retrieved from the long-term memory and training cannot proceed. Henceforth, proper learning of earlier experiences is fundamental for the development of higher-complexity tasks.
The human-system interface will produce the following outcome:

*TEA: Nao take the teabag
*SYS: _teabag _grab _lift
*TEA: Nao take the bottle
*SYS: _bottle _grasp _lift

The robot executes both actions in sequence with the referent object. The human can continue to build increasingly complex tasks with further higher-order instructions, whilst recalling the outcomes of the earlier-learned tasks in the learning of new tasks.

Below are illustrated further higher-order steps from the snowballing toolchain.

(training, 2nd higher order)

*TEA: place the mug on the table
*SYS: _mug _grasp _pick-up _table _drop

(test, 2nd higher order)

*TEA: add the water to the mug
*SYS: _bottle _grasp _lift _mug _pour

*TEA: add the teabag to the mug
*SYS: _teabag _grab _lift _mug _throw

(training, 3rd higher order)

*TEA: Nao make me tea
*SYS: _mug _grasp _pick-up _table _drop

(test, 3rd higher order)

*TEA: add the water to the mug
*SYS: _bottle _grasp _lift _mug _pour

*TEA: add the teabag to the mug
*SYS: _teabag _grab _lift _mug _throw

To build novel tasks (e.g., make me a juice, make me a toast), the user must only repeat the tests in sequence (test – primitive actions, test – 1st higher-order, test – 2nd higher-order, test 3rd higher-order) using the target objects and actions. No further training is required. The learning framework can abstract and generalise accordingly by following the same line of reasoning (see Figure 23, main text).

---

8 This task re-uses the earlier-learned experience of procuring the mug (_mug _grasp _pick-up) without training it anew.

9 This task re-uses the earlier-learned experience of placing the mug on the table (_mug _grasp _pick-up _table _drop) and those that are self-memorised during the runtime exploitation (test) of adding water to the cup (_bottle _grasp _lift _cup _pour) and adding the teabag to the cup (_teabag _grab _lift _mug _throw), without training them anew.
Bibliography


131

Bibliography


Google Cloud Platform Services (APIs) [online], available at https://cloud.google.com/apis/docs/overview, last accessed February 2022


Hinaut, X. and Wermer, S., 2014, September. An incremental approach to language acquisition: Thematic role assignment with echo state networks. In *International Conference on Artificial Neural Networks* (pp. 33-40). Springer, Cham. [https://doi.org/10.1007/978-3-319-11179-7_5](https://doi.org/10.1007/978-3-319-11179-7_5).

Hinaut, X., Twiefel, J., Petit, M., Dominey, P. and Wermer, S., 2015, December. A recurrent neural network for multiple language acquisition: Starting with english and french. In *Proceedings of the NIPS Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (CoCo 2015)*. [https://hal.inria.fr/hal-02561258](https://hal.inria.fr/hal-02561258).


iCub humanoid robot, Italian Institute of Technology (IIT) [online], available at [https://icub.iit.it/](https://icub.iit.it/), last accessed February 2022.


Jawahar, G., Sagot, B. and Seddah, D., 2019, July. What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*. [https://hal.inria.fr/hal-02131630](https://hal.inria.fr/hal-02131630).


McNab, F. and Klingberg, T., 2008. Prefrontal cortex and basal ganglia control access to working memory. *Nature neuroscience, 11*(1), pp.103-107 [https://doi.org/10.1038/nn2024](https://doi.org/10.1038/nn2024)


Mirolli, M. and Parisi, D., 2006. Talking to oneself as a selective pressure for the emergence of language. In *The evolution of language* (pp. 214-221) [https://doi.org/10.1142/9789812774262_0028](https://doi.org/10.1142/9789812774262_0028)


