ISSN 2427-4577

BULLETIN N° 221
ACADÉMIE EUROPEENNE INTERDISCIPLINAIRE DES SCIENCES

INTERDISCIPLINARY EUROPEAN ACADEMY OF SCIENCES

Lundi 8 janvier 2018:

à 15h45 à l'Institut Henri Poincaré, 11 rue Pierre et Marie Curie 75005 PARIS

Conférence de Marie AMALRIC ,
Chercheuse Post Doc au Département du Cerveau des Sciences Cognitives
Lab CAos/ Université de Rochester / Etat de New York/USA
"Comment le cerveau humain manipule-t-il les concepts mathématiques ?"

Notre Prochaine séance aura lieu le lundi 5 février 2018 à 15h45
à l'Institut Henri Poincaré salle 05
11, rue Pierre et Marie Curie 75005 PARIS

Elle aura pour thème

Conférence d'Emmanuel DUPOUX,
Directeur d'Etudes 1 à Ecole des Hautes Etudes en Sciences Sociales (EHESS)
Laboratoire des Sciences cognitives et Psycholinguistique
ENS-EHESS-CNRS-INRIA 29 rue d'Ulm 75005 PARIS
" L'Intelligence Artificielle peut elle aider l'étude du développement cognitif ? (et vice versa)?"

Académie Européenne Interdisciplinaire des Sciences
Siège Social : 5 rue Descartes 75005 Paris
http://www.science-inter.com

ACADÉMIE EUROPÉENNE INTERDISCIPLINAIRE DES SCIENCES
INTERDISCIPLINARY EUROPEAN ACADEMY OF SCIENCES

janvier 2018

N°221

TABLE DES MATIERES

<u>Prochaine séance : lundi 8 janvier 2018</u>

Conférence d'Emmanuel DUPOUX,
Directeur d'Etudes 1 à Ecole des Hautes Etudes en Sciences Sociales (EHESS)
Laboratoire des Sciences cognitives et Psycholinguistique
ENS-EHESS-CNRS-INRIA 29 rue d'Ulm 75005 PARIS
" L'Intelligence Artificielle peut elle aider l'étude du développement cognitif ? (et vice versa)?"

Académie Européenne Interdisciplinaire des Sciences
Siège Social : 5 rue Descartes 75005 Paris
http://www.science-inter.com

ACADEMIE EUROPEENNE INTERDISCIPLINAIRE DES SCIENCES
INTERDISCIPLINARY EUROPEAN ACADEMY OF SCIENCES
5 rue Descartes 75005 PARIS

Séance du Lundi 8 janvier 2018/Institut Henri Poincaré à 15h45

La séance est ouverte à 15h45 **sous la Présidence de Victor MASTRANGELO** et en la présence de nos Collègues Gilbert BELAUBRE, Jean-Louis BOBIN, Gilles COHEN-TANNOUDJI, Françoise DUTHEIL , Claude ELBAZ, Jean -Pierre FRANÇOISE, Michel GONDRAN, Irène HERPE-LITWIN, Gérard LEVY, Claude MAURY, Marie-Françoise PASSINI, Jacques PRINTZ, Jean SCHMETS , Jean-Pierre TREUIL, Alain STAHL,. Jean-Paul TEYSSANDIER .

Etaient présents en tant que visiteurs Jean BERBINAU administrateur du lycée Saint Louis et du Collège Stanislas, Marie-Joséphe MARTIN ancienne professeur de Mathématiques en classes préparatoires..

Etaient excusés :François BEGON, Jean-Pierre BESSIS, Jean-Louis BOBIN, Bruno BLONDEL, Michel CABANAC, Alain CARDON, Juan-Carlos CHACHQUES, Alain CORDIER , Daniel COURGEAU, Sylvie DERENNE, Ernesto DI MAURO, Jean-Felix DURASTANTI, Claude ELBAZ, Vincent FLEURY, Robert FRANCK, Dominique LAMBERT, Valérie LEFEVRE-SEGUIN, Antoine LONG, Pierre MARCHAIS, Anastassios METAXAS, Jean-Jacques NIO, Alberto OLIVIERO, Edith PERRIER, Pierre PESQUIES, Michel SPIRO, Mohand TAZEROUT , Jean VERDETTI.

I. . Présentation de notre conférencière Marie AMALRIC par Victor MASTRANGELO :

Marie AMALRIC, née en 1989, nous a été recommandée par le Pr Stanislas DEHAENE lors de son intervention du 11 septembre 2017. Elle est Post -doctorante en Sciences Cognitives à l'Université de ROCHESTER dans l'Etat de new York aux USA dans le département "Cerveau et Sciences cognitives" . Elle a travaillé notamment en partenariat avec le Pr Stanislas DEHAENE(INSERM-CEA / Collège de France) sur l'Etude par (fMRI) de l'acquisition de concepts mathématiques de haut niveau chez les étudiants
Un langage cérébral pour les formes géométriques.

ETUDES:

2013-2017	**Université Pierre et Marie Curie (UPMC)** **Doctorat en Neurosciences Cognitive, sous la direction de Stanislas Dehaene**. “***Etude des mécanismes cérébraux impliqués dans l'apprentissage et le traitement des concepts mathématiques de haut niveau***"
2009-2013	**École Normale Supérieure (ENS Paris, Ulm)**
2010-2013	**École Nationale Supérieure des Techniques Avancées** (ENSTA Paris-tech), **Master Ingénierie Mathématique** Specialité: ***Optimisation, recherche opérationnelle et commande*** .
2010-2012	**École Normale Supérieure: Masters Sciences cognitives**

	Sujets Master: -2010: *Origines des idées mathematiques.* (Dir. Giuseppe Longo, CREA). -2011: '*Modélisation Bayésienne des comportements des rats dans un labyrinthe ’* (Dir. J. Droulez LPPA, Collège de France). -2012: *Intuition et traitement cortical des concepts mathématiques de haut niveau.* (Dir. Stanislas Dehaene, UNICOG).
2009-2010	**Université Pierre et Marie Curie (UPMC), Paris 6,** **Licence de Mathematiques** Principal: Algebra secondaire: Sciences Cognitive àl'ENS Stage : *Modelisation des explosions de neurones*. (Dir. Sophie Denève, LNC, ENS)
2007-2009	**Classe Préparatoire aux Grandes Ecoles “Grandes Écoles” en Mathematiques and Physique au Lycée Louis le Grand (Paris)**

CARRIERE DE RECHERCHE

2017-… : University of Rochester, Department of Brain and Cognitive Sciences, CAOs lab, Post-doc position.

2016-2017: Collège de France, ATER.

2013-2016: UPMC, INSERM-CEA Unité de Neuro-imagerie Cognitive, recherche doctorale

PRINCIPALES PUBLICATIONS:

Amalric, M. & Dehaene, S. *Origins of the brain networks for advanced mathematics in expert mathematicians.* PNAS, 04/2016; 113(18). DOI:10.1073/pnas.1603205113

Amalric, M., Wang, L., Pica, P., Figueira, S., Sigman, M., Dehaene, S. *The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers.* PLOS Computational Biology, 01/2017. DOI:10.1371/journal.pcbi.1005273

Amalric, M., Denghien, I., Dehaene, S. *On the role of visual experience in mathematical development: Evidence from blind mathematicians.* Developmental Cognitive Neuroscience, 10/2017. DOI:10.1016/j.dcn.2017.09.007

Amalric, M., Dehaene, S. *Cortical Circuits for Mathematical Knowledge: Evidence for a Major Subdivision within the Brain Semantic Networks.* Phil Trans Royal Society B, 12/2017. DOI:10.1098/rstb.2016.0515

COMMUNICATIONS DANS DES SEMINAIRES OU DEPARTEMENTS DE RECHERCHE

Mai 2017: Université du Luxembourg, invitée par Christine Schiltz.

Janvier 2017: Université de Gand , invitée par Wim Fias.

Novembre 2016: MIT, Cambridge, invitée par Josh Tenenbaum.

Octobre 2016: University du Wisconsin, Madison, invitée par Edward Hubbard.

Octobre 2016: University de Pennsylvanie, Philadelphie, invitée par Elizabeth Brannon.

Novembre 2015: LPP, Paris, invitée par Véronique Izard.

Mars 2015: PICNIC Lab, ICM, Paris, invitée par Imen el Karoui. e the lab FELLOWSHIPS AND AWW

BOURSES et RECOMPENSES

2013: **Récompense du meilleur stage final**, des anciens de l'ENSTA-ParisTech pour le projet de recherche “ Etude des processus cérébraux d'apprentissage des concepts mathématiques abstraits "

2013: **Bourse doctorale**, du “DIM-Cerveau et pensée” (Région Ile-de-France).

2016: Bourse de voyage de l'ISC pour participer à une université d'été sur le raisonnement à l' UQAM, Montréal.

2017: Bourse post-doctorale, de la fondation Fyssen, pour le projet: “Acquisition de Concepts Mathématiques dans le Cerveau Humain ”.

II. Conférence de Marie AMALRIC

Résumé de la conférence avec références bibliographiques:

Comment le cerveau humain manipule-t-il les concepts mathématiques ?

Comment le cerveau humain parvient-il à conceptualiser des idées abstraites ? Quelle est en particulier l'origine de l'activité mathématique lorsqu'elle est associée à un haut niveau d'abstraction ? Cette question qui intéresse depuis longtemps philosophes, mathématiciens et enseignants, commence aujourd'hui à être abordée par les neurosciences cognitives, et ce, en grande partie par le biais d'études portant sur l'arithmétique élémentaire. Toutefois, les mathématiques recouvrent de nombreuses disciplines telles que l'algèbre, l'analyse ou la géométrie et ne sauraient être réduites à la compréhension des nombres. Aussi mon travail privilégie l'étude de la manipulation d'idées mathématiques plus avancées en cherchant à identifier les corrélats neuronaux de la réflexion mathématique de haut niveau.

Dans son exposé, elle présentera les résultats de trois expériences en IRMf (Imagerie par Résonance Magnétique fonctionnelle) , menées chez des mathématiciens professionnels (dont trois mathématiciens non-voyants) qui devaient évaluer la valeur de vérité d'affirmations mathématiques et non-mathématiques énoncées oralement. Même formulées comme des phrases, toutes les affirmations mathématiques, quels que soient leur difficulté , leur domaine, ou l'expérience visuelle du participant, impliquent systématiquement des régions cérébrales totalement dissociées des aires reliées au langage et au traitement sémantique, mais qui coïncident avec des zones activées par l'arithmétique élémentaire. A l'inverse, même lorsqu'elles comprennent des opérateurs logiques (quantificateurs, négation), les affirmations non-mathématiques (portant sur l'histoire, les arts, la géologie, la faune etc…), activent aires cérébrales classiquement associées au langage. L'activité mathématique semble donc « recycler » des aires cérébrales impliquées dans la connaissance élémentaire des nombres et de l'espace et se dissocier de la manipulation sémantique du langage.

References:
Amalric, M. &amp; Dehaene, S. Origins of the brain networks for advanced mathematics in expert mathematicians. PNAS, 04/2016; 113(18). DOI:10.1073/pnas.1603205113
Amalric, M., Denghien, I., Dehaene, S. On the role of visual experience in mathematical development: Evidence from blind mathematicians. Developmental Cognitive Neuroscience, 10/2017. DOI:10.1016/j.dcn.2017.09.007

Un compte-rendu détaillé sera prochainement disponible sur le site de l'AEIS , http://www.science-inter.com

Annonces

I. Le prochain colloque de l'AEIS sur "Les Signatures de la Conscience" se tiendra les jeudi 15 mars et vendredi 16 mars 2018 à l'Institut Henri Poincaré dans l'Amphi Hermite . Pour vous inscrire il vous suffit d'aller sur le site :

https://aeis-2018.sciencesconf.org

II. Notre Collègue Alain STAHL vient de publier auprès de la Librairie Philosophique VRIN la 3ème édition de son ouvrage "*Science et Philosophie*"

Cet ouvrage de 337 pages est consacré à une réflexion sue les conséquences épistémologiques et philosophiques des avancées spectaculaires dans tous les domaines scientifiques. Il renvoie à d'importants développements donnés en libre accès sur le site de l'auteur http://perso.wanadood.fr/alain.stahl

Les apports nouveaux, dans cette troisième édition, concernent :

1 - des acquis récents qui étayent ses réflexions de « critique scientifique » sur des points d'actualité, tels que le calcul informatique, les transitions de phase, la cosmologie, le repliement des protéines, l'intelligence artificielle, les méthodes de mesure...
2 - Un dernier chapitre, entièrement nouveau, où – par une méthode originale, récapitulant les conclusions des chapitres scientifiques – l'auteur tente de répondre à la question posée par le nouveau sous-titre de l'ouvrage. : "La science permet-elle une présentation moderne des grandes questions philosophiques?"
L'écriture est rigoureuse, mais la lecture est aisée.

Les grands thèmes philosophiques sont toujours, –chose rare -, étayés par la priorité donnée aux acquis scientifiques. C'est une mise à niveau dont la lecture induit un dialogue permanent, très ouvert et très riche, avec l'auteur.

III. **Quelques ouvrages papiers relatifs au colloque de 2014 " Systèmes stellaires et planétaires-Conditions d'apparition de la Vie"** -

- Prix de l'ouvrage :25€ .
- Pour toute commande s'adresser à :

Irène HERPE-LITWIN Secrétaire générale AEIS
39 rue Michel Ange 75016 PARIS
06 07 73 69 75
irene.herpe@science-inter.com

IV. **L'ouvrage cité ci-dessus est accessible gratuitement sur le site**:

http://www.edp-open.org/images/stories/books/fulldl/Formation-des-systemes-stellaires-et-planetaires.pdf

Documents

Pour compléter l'intervention de Marie AMALRIC qui a cité les travaux du mathématicien Grothendieck, notre collègue Jacques PRINTZ nous a confié une petite fiche relative à ce grand mathématicien:

p. 08 : Fiche Grothendieck écrite par notre Collègue Jacques PRINTZ

Pour préparer l'intervention de notre conférencier Emmanuel DUPOUX sur les relations entre linguistique et Intelligence Artificielle nous vous proposons :

p.09 : issu du site https://arxiv.org/abs/1607.08723v3 un article d'Emmanuel DUPOUX de 2016 intitulé :" Cognitive Science in the era of Artificial Intelligence: ***A roadmap for reverse-engineering the infant language-learner."***

P. 34 : issu du site http://www.lscp.net/persons/dupoux/papers/Linzen_DG_2017_Assessing%20syntax-sensitive%20dependencies%20in%20LSTMs.TACL.pdf un article de 2017 de Linzen, T., Dupoux, E. & Goldberg, Y. (2016). "***Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Transactions of the Association for Computational Linguistics, 4, 521-535***".

Fiche Grothendieck par notre Collègue Jacques PRINTZ

Dans l'autobiographie d'Alexandre Grothendieck, décédé en 2014, et refondateur de la Géométrie algébrique moderne[1], page 48 de *Récoltes et semailles* [N'est disponible que sur Internet, téléchargeable sans problème, sauf que ça fait plus de 1.000 pages ; https://www.quarante-deux.org/archives/klein/prefaces/Romans_1965-1969/Recoltes_et_semailles.pdf], il y a un petit texte tout à fait illustratif de sa démarche architecturale, en rapport avec ce que nous a expliqué Marie Amalric, dans sa conférence du 8/01/2018 :

« La structure d'une chose n'est nullement une chose que nous puissions "inventer". Nous pouvons seulement la mettre à jour patiemment, humblement en faire connaissance, la "découvrir". S'il y a inventivité dans ce travail, et s'il nous arrive de faire oeuvre de forgeron ou d'infatigable bâtisseur, ce n'est nullement pour "façonner", ou pour "bâtir", des "structures". Celles-ci ne nous ont nullement attendues pour être, et pour être exactement ce qu'elles sont ! Mais c'est pour exprimer, le plus fidèlement que nous le pouvons, ces choses que nous sommes en train de découvrir et de sonder, et cette structure réticente à se livrer, que nous essayons à tâtons, et par un langage encore balbutiant peut-être, à cerner. Ainsi sommes-nous amenés à constamment "inventer" le langage apte à exprimer de plus en plus finement la structure intime de la chose mathématique, et à "construire" à l'aide de ce langage, au fur et à mesure et de toutes pièces, les "théories" qui sont censées rendre compte de ce qui a été appréhendé et vu. Il y a là un mouvement de va-et-vient continuel, ininterrompu, entre l'appréhension des choses, et l'expression de ce qui est appréhendé, par un langage qui s'affine et se re-crée au fil du travail, sous la constante pression du besoin immédiat ».

Cette approche « linguistique », de type grammaire, n'est pas la première du genre, von Neumann a tenu des propos similaires dans son dernier livre *The computer and the brain*, écrit quelques mois avant sa mort, en 1957 ; et également chez Turing dans son article de 1936, en réponse aux problématiques introduites par Hilbert et Russel/Whitehead avec le langage des *Principia mathematica*. En informatique, ce genre de démarche est fondamentale, je l'avais brièvement abordée dans la présentation de mes travaux lors de notre séance du 9 mai 2017, où l'on peut représenter l'architecture des systèmes par une cascade de langages emboîtés les uns dans les autres, ceux utilisés par les programmeurs n'ayant qu'un très lointain rapport avec ceux utilisés pour piloter les machines de gravure pour « sculpter » les cristaux de silicium qui in fine sont les transducteurs énergétiques qui réalisent concrètement ce que le programmeur a prescrit.

Plus près de nous, il y a également le livre du mathématicien Jean-Marie Souriau, *Grammaire de la nature*, spécialiste des systèmes dynamiques, également très intéressant : [http://www.jmsouriau.com/Publications/Grammaire%20de%20la%20Nature/JMSouriau-GrammaireDeLaNature8juillet2007-complet.pdf].

Pour conclure, l'approche présentée par M. Almaric, intéressante, me paraît un peu réductionniste, car si on essaye de comprendre ce qui se passe dans un ordinateur en scrutant les transistors et le silicium, on est sûr de ne rien comprendre, en tout cas pas ce que font les programmeurs ...

[1] Le tome 1 du cours de Jean Dieudonné, *Cours de géométrie algébrique*, est une excellente introduction historique, des origines grecques jusqu'à Grothendieck.

Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner

Emmanuel Dupoux
EHESS, ENS, PSL Research University, LSCP, CNRS
emmanuel.dupoux@gmail.com, www.syntheticlearner.net

Abstract

During their first years of life, infants learn the language(s) of their environment at an amazing speed despite large cross cultural variations in amount and complexity of the available language input. Understanding this simple fact still escapes current cognitive and linguistic theories. Recently, spectacular progress in the engineering science, notably, machine learning and wearable technology, offer the promise of revolutionizing the study of cognitive development. Machine learning offers powerful statistical learning algorithms that can achieve human-like performance on many linguistic tasks. Wearable sensors can capture vast amounts of data, which enable the reconstruction of the sensory experience of infants in their natural environment. The project of 'reverse engineering' language development, i.e., of building an effective system that mimics infant's achievements appears therefore to be within reach.

Here, we analyze the conditions under which such a project can contribute to our scientific understanding of early language development. We argue that instead of defining a sub-problem or simplifying the data, computational models should address the full complexity of the learning situation, and take as input as faithful reconstructions of the sensory signals available to infants as possible. This implies that accessible but privacy-preserving repositories of home data be setup and widely shared, and models be evaluated at different linguistic levels through a benchmark of psycholinguist tests that can be passed by machines and humans alike, linguistically and psychologically plausible learning mechanisms be merged with probabilistic/optimization principles from machine learning to yield scalable learning architectures. We discuss the feasibility of this approach and present preliminary results.

Keywords

Artificial intelligence, psycholinguistics, computational modeling, corpus analysis, early language acquisition, infant development, language bootstrapping, machine learning, phonetic learning

1 Introduction

In recent years, Artificial Intelligence (AI) has been hitting the headlines with impressive achievements at matching or even beating humans in complex cognitive tasks (playing go or video games: Mnih et al., 2015; Silver et al., 2016; processing speech and natural language: Amodei et al., 2015a; Ferrucci, 2012; recognizing objects and faces: He, Zhang, Ren, & Sun, 2015; Lu & Tang, 2014) and promising a revolution in manufacturing processes and human society at large. These successes show that with statistical learning techniques, powerful computers and large amounts of data, it is possible to mimic important components of human cognition. What does it tell us about the underlying psychological and/or neural processes that are used by humans to solve these tasks? Can AI also revolutionize the study of human cognition by providing us with scientific insights about these processes? Here, we argue that developmental psychology and in particular, the study of language acquisition is one area where, indeed, AI and machine learning advances can be transformational, provided that the involved fields make significant adjustments in their practices in order to adopt what we call the *reverse engineering approach*. Specifically:

> The reverse engineering approach to the study of infant language acquisition consists in constructing computational systems that can, when fed with the same input data, reproduce language acquisition as it is observed in infants.

The idea of using machine learning or AI techniques as a means to study child's language learning is actually not new (to name a few: Kelley, 1967; Anderson, 1975; Berwick, 1985; Rumelhart & McClelland, 1987; Langley & Carbonell, 1987) although relatively few studies have concentrated on the early phases of language learning (see Brent, 1996b, for a review). What is new, however, is that whereas previous AI approaches were limited to proofs of principle on toy or miniature languages, modern AI techniques have scaled up so much that end-to-end language processing systems working with real inputs are now deployed commercially. This paper examines whether and how such unprecedented change in scale could be put to use to address lingering scientific questions in the field of language development. The structure of the paper is as follows: In Section 2, we present two deep scientific puzzles that large scale modeling approaches could in principle address: solving the bootstrapping problem, accounting for developmental trajectories. In Section 3, we review past theoretical or modeling work, showing that these puzzles have not, so far, received an adequate answer. In Section 4, we argue that to answer them with reverse engineering, four requirements have to be addressed: (1) modeling should be done on real data, (2) model performance should be compared with that of humans, (3) modeling should be computationally effective, and (4) datasets, benchmarks and models should be open sourced. In Section 5, we argue that within a simplifying framework, these requirements can be reached given current technology, although specific roadblocks need to be lifted. In Section 6 we show that even before these roadblocks are lifted, interesting results can be obtained. In Section 7 we show how the reverse engineering approach can be generalized beyond the simplifying framework presented in Section 5, and we conclude in Section 8.

2 Two scientific puzzles of early language development

Language development is a theoretically important subfield within the study of human cognitive development for the following reasons: First, the linguistic system is uniquely *complex*: mastering a language implies mastering a combinatorial sound system (phonetics and phonology), an open ended morphologically structured lexicon, and a compositional syntax and semantics (e.g., Jackendoff, 1997). No other animal communication system uses such a complex multilayered organization. On this basis, it has been claimed that humans have evolved (or acquired through a mutation) an innately specified computational architecture to process language (see Chomsky, 1965; Hauser, Chomsky, & Fitch, 2002; Steedman, 2014). Second, the overt manifestations of this system are extremely *variable* across languages and cultures. Language can be expressed through the oral or manual modality. In the oral modality, some languages use only 3 vowels, other more than 20. Consonants inventories vary from 6 to more than 100. Words can be mostly composed of a single syllable (as in Chinese) or long strings of stems and affixes (as in Turkish). Semantic roles can be identified through fixed positions within constituents, or be identified through functional morphemes, etc. (see Song, 2010, for a typology of language variation). Evidently, infants acquire the relevant variant through learning, not genetic transmission. Third, the human language capacity can be viewed as a finite computational system with the ability to generate a (virtual) infinity of utterances. This turns into a *learnability problem* for infants: on the basis of finite evidence, they have to induce the (virtual) infinity corresponding to their language. As has been discussed since Aristotle, such induction problems do not have a generally valid solution. Therefore, language is simultaneously a human-specific biological trait, a highly variable cultural production, and an apparently intractable learning problem. Despite these complexities, most infants spontaneously learn their native(s) language(s) in a matter of a few years of immersion in a linguistic environment. The more we know about this simple fact, the more puzzling it appears. Specifically, we outline two central scientific puzzles that a reverse engineering approach could, in principle help to solve: solving the bootstrapping problem and accounting for developmental trajectories. The first puzzle relates to the ultimate outcome of language learning: the so-called *stable state*, i.e., the language competence in the idealized adult. The second puzzle relates to what we know of the intermediate steps in the acquisition process, and their variations as a function of language input.[1]

2.1 Solving the bootstrapping problem

The stable state can be described as the operational knowledge (which we will refer to here broadly as a 'grammar') which enables adults to process a virtual infinity of utterances in their native language. This grammar is a multilayered system comprising several components: phonetics, phonology, morphology, syntax, semantics, pragmatics. The *bootstrapping problem* arises from the fact these different components appear *interdependent* from a learning point of view. For instance, the phoneme inventory of a language is defined through pairs of words that differ minimally in sounds (e.g., "light" vs "right"). This would suggest that to learn phonemes, infants need to first learn words. However, from a processing viewpoint, words are recognized through their phonological constituents (e.g., Cutler, 2012), suggesting that infants should learn phonemes before words. Similar paradoxical co-dependency issues have been noted between other linguistic levels (for instance, syntax and semantics: Pinker, 1987, prosody and syntax: Morgan & Demuth, 1996). In other words, order to learn any one component of the language competence, many others need to be learned first, creating apparent circularities. The bootstrapping problem is further compounded by the fact that infants do not have to be taught formal linguistics or language courses to learn their native language(s). As in other cases of animal communication, infants *spontaneously* acquire the language(s) of their community by merely being immersed in that community (Pinker, 1994). Experimental and observational studies have revealed that infants start acquiring elements of their language (phonetics, phonology, lexicon, syntax and semantics) even before they can talk (Jusczyk, 1997; Hollich et al., 2000; Werker & Curtin, 2005), and therefore before parents can give them much feedback about

[1] The two puzzles are not independent as they are two facets of the same phenomenon. In practice, proposals for solving the bootstrapping problem may offer insights about the observed trajectories. Vice-versa, data on developmental trajectories may provide more manageable subgoals for the difficult task of solving the bootstrapping problem.

their progress into language learning. This suggests that language learning (at least the initial bootstrapping steps) occurs largely *without supervisory feedback*.[2] A reverse engineering approach has the potential of solving this puzzle by providing a system that can demonstrably bootstrap into language when fed with similar, supervisory poor, inputs.

2.2 Accounting for developmental trajectories

In the last forty years, a large body of empirical work has been collected regarding infant's language achievements during their first years of life. This work has only added more puzzlement. First, given the multi-layered structure of language, one could expect a stage-like developmental tableau where acquisition would proceed as a discrete succession of learning phases organized logically or hierarchically (e.g., building linguistic structure from the low level to the high levels). This is not what is observed (see Figure 1). For instance, infants start differentiating native from foreign consonants and vowels at 6 months, but continue to fine tune their phonetic categories well after the first year of life (e.g., Sundara, Polka, & Genesee, 2006). However, they start learning about the sequential structure of phonemes (phonotactics, see Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993) way *before* they are done acquiring the phoneme inventory (Werker & Tees, 1984). Even before that, they start acquiring the meaning of a small set of common words (e.g. Bergelson & Swingley, 2012). In other words, instead of a stage-like developmental tableau, the evidence shows that acquisition takes places at all levels more or less simultaneously, in a *gradual* and largely *overlapping* fashion. Second, observational studies have revealed considerable *variations* in the *amount of language input* to infants across cultures (Shneidman & Goldin-Meadow, 2012) and across socio-economic strata (Hart & Risley, 1995), some of which can exceed an *order of magnitude* (Weisleder & Fernald, 2013, p. 2146). These variations do impact language achievement as measured by vocabulary size and syntactic complexity (Hoff, 2003; Huttenlocher, Waterfall, Vasilyeva, Vevea, & Hedges, 2010; Pan, Rowe, Singer, & Snow, 2005; Rowe & Goldin-Meadow, 2009, among others), but at least for some markers of language achievement, the differences in outcome are much less extreme than the variations in input. For canonical babbling, for instance, an order of magnitude would mean that some children start to babble at 6 months, and others at 5 years! The observed range is between 6 and 10 months, less than a 1 to 2 ratio. Similarly, reduced range of variations are found for the onset of word production and the onset of word combinations. This suggests a surprising level of *resilience* to language learning, i.e., some minimal amount of input is sufficient to trigger certain landmarks. A reverse engineering approach has the potential of accounting for this otherwise perplexing developmental tableau, and provide quantitative predictions both across linguistic levels (gradual overlapping pattern), and cultural or individual variations in input (resilience).

2.3 Why it matters

Solving these two puzzles would have a large impact on the study of language learning and more broadly on developmental psychology. At the theoretical level, it would enable give closure to long-lasting controversies (nature versus nurture, semantic versus syntactic bootstrapping) and replace them with quantitative evaluations of the relative contribution of each types of factors. In other words, it would transform a field which is predominantly descriptive and some would say speculative, into a field where formal models yield quantitative predictions. It would also give rise to practical applications in the fields of developmental disorders and language education by enabling predictive models of learning trajectories as a function of input in naturalistic environement or in interventional studies. Finally, it could have an impact in the field of AI itself through the establishment of learning architectures able to learn linguistic structures much more autonomously and robustly than is currently achieved.

3 Past work

Early language acquisition is primarily an empirical field of research. Much of what we know has been obtained thanks to the patient accumulation of data in two lines of work. The first one is devoted to the collection and manual transcription of parents and infants interactions. A large number of datasets across languages have been collected and organized into repositories that have proved immensely useful to the research community. One prominent example of this is the CHILDES repository (MacWhinney, 2000), which has enabled more than 5000 research papers (according to a google scholar search as of 2016). The other line consists in measuring the linguistic knowledge of infants of various ages across different languages through the administration of experimental tests (see Jusczyk, 1997; Bornstein & Tamis-LeMonda, 2010 for reviews). Besides this impressive activity in data gathering, the field of language development is also actively pursuing theoretical work. Here, we briefly review three major strands related to psycholinguistics, formal linguistics and AI, respectively, and argue that even though these strands have provided important insights into the acquisition process, they still fall short of accounting for the two puzzles presented in Section 2.

[2] Even in later acquisitions, the nature, universality and effectiveness of corrective feedback of children's outputs has been debated (see Brown, 1973; Pinker, 1989; Marcus, 1993; Chouinard & Clark, 2003; Saxton, 1997; Clark & Lappin, 2011).

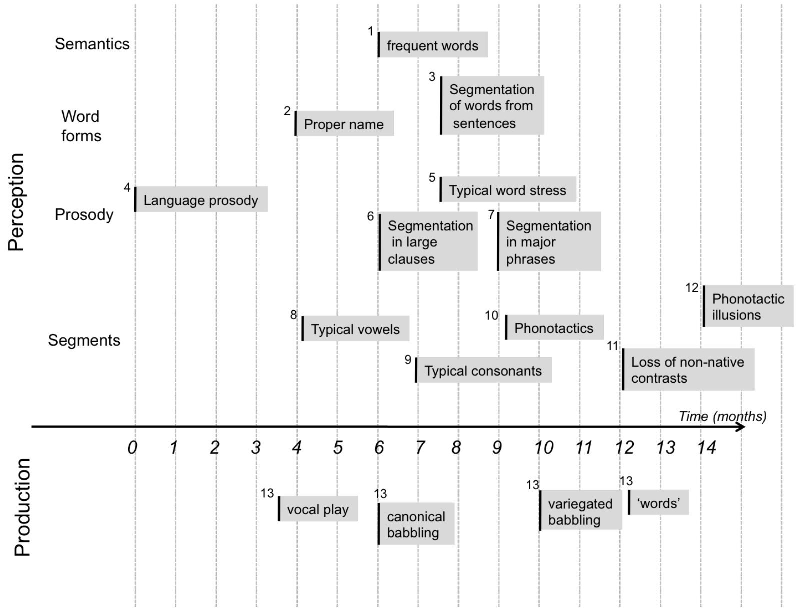

Figure 1. Sample studies illustrating infant's language development. The left edge of each box is aligned to the earliest age at which the result has been documented. [1] Tincoff & Jusczyk, (1999); Bergelson & Swingley, (2012); [2] Mandel et al. (1995); [3] Jusczyk & Aslin (1995) [4] Mehler et al. (1988) [5] Jusczyk et al. (1999) [6] Hirsh-Pasek et al. (1987) [7] Jusczyk et al (1992) [8] Kuhl et al. (1992) [9] Eilers et al. (1979) [10] Jusczyk et al. (1993) [11] Werker & Tees (1984) [12] Mazuka et al. (2011) [13] Stark (1980).

3.1 Conceptual frameworks and learning mechanisms

Within developmental psycholinguistics, *conceptual frameworks* have been proposed to account for key aspects of the developmental trajectories (the competition model: Bates & MacWhinney, 1987; MacWhinney, 1987 ; WRAPSA: Jusczyk, 1997; the emergentist coalition model: Hollich et al., 2000; PRIMIR: Werker & Curtin, 2005; the usage-based theory: Tomasello, 2003; among others). These frameworks present overarching architectures or scenarios that integrate many empirical results. WRAPSA (Jusczyk, 1997) focuses on phonetic learning and lexical segmentation during the first year of life. PRIMIR (Werker & Curtin, 2005) extends WRAPSA by incorporating phonetic and speaker-related categories at an early stage, and meaning and phonemic categories at a later stage. The emergentist coalition model (Hollich et al., 2000) focuses on the attentional, social and linguistic factors that modulate the association between lexical forms and meanings at different ages. The competition model (Bates & MacWhinney, 1987; MacWhinney, 1987) and the usage-based theory (Tomasello, 2003) focus on grammar learning; the former is lexicon-based and focuses on mechanisms of competitive learning. The latter is construction-based and focuses on social and pragmatic learning mechanisms. While these conceptual framework are very useful in summarizing and organizing a vast amount of empirical results, and could serve as sources of inspiration for computational models, they are not specific enough to address our two scientific puzzles. They tend to refer to mechanisms using verbal descriptions (*statistical learning*, *rule learning*, *abstraction*, *grammaticalization*, *analogy*) or boxes and arrows diagrams. This type of presentation may be intuitive, but also vague. The same description may correspond to many different computational mechanisms which would yield different predictions. These frameworks are therefore difficult to put to empirical test. In addition, because they are not formal, one cannot demonstrate that these models can effectively solve the language bootstrapping problem. Nor do they provide quantitative predictions about the observed resilience in developmental trajectories or their variations as a function of language input at the individual, linguistic or cultural level. Psycholinguists sometimes supplement conceptual frameworks with propositions for specific *learning mechanisms* which are tested using an artificial language paradigm. As an example, a mechanism based on the tracking of statistical modes in phonetic space has been proposed to underpin phonetic category learning in infancy. It was tested in infants through the presentation of a simplified language (a continuum of syllables between /da/ and /ta/) where the statistical distribution of acoustic tokens was controlled (Maye, Werker, & Gerken, 2002). It was also modeled computationally using unsupervised clustering algorithms and tested using simplified corpora or synthetic data (Vallabha, McClelland, Pons, Werker, & Amano, 2007; McMurray, Aslin, & Toscano, 2009). A similar double-pronged approach (experimental and modeling evidence) has been conducted for other mechanisms: word segmentation based on transition probability (Saffran, Aslin, & Newport, 1996; Daland & Pierrehumbert, 2011), word meaning learning based on cross situational statistics (Yu & Smith, 2007; K. Smith, Smith, & Blythe, 2011; Siskind, 1996), semantic role learning based on syntactic cues (Connor, Fisher, & Roth, 2013), etc. Although studies with artificial languages are useful to discover candidate learning algorithms which could be incorporated in a global architecture, the algorithms proposed have only been tested on toy or artificial languages; there is therefore no guarantee that they would actually work when faced with realistic corpora that are both very large and very noisy. In fact, as discussed in section 6.1, some of these algorithms do not scale up. In addition, it remains to be shown that taken collectively, such learning mechanisms (or scaled up versions thereof) would work synergistically to solve the bootstrapping problem, as opposed to cancelling each other's out.

3.2 Formal linguistic models

Even though much of current theoretical linguistics is devoted to the study of the language competence in the stable state, very interesting work has also be conducted in the area of formal models of *grammar induction*. These models propose algorithms that are provably powerful enough to learn a fragment of grammar given certain assumptions about the input. For instance, Tesar and Smolensky (1998) proposed an algorithm that provided pairs of surface and underlying word forms can learn the phonological grammar (see also Magri, 2015). Similar learnability assumptions and results have been obtained for stress systems (Dresher & Kaye, 1990; Tesar & Smolensky, 2000). For learnability results of syntax, see the review in Clark and Lappin (2011). These models establish important learnability results, and in particular, demonstrate that under certain hypotheses, a particular class of grammar is learnable. What they do not demonstrate however is that these hypotheses are met for infants. In particular, most grammar induction studies assume that infants have an error-free, adult-like symbolic representation of linguistic entities (e.g., phonemes, phonological features, grammatical categories, etc). Yet, perception is certainly not error-free, and it is not clear that infants have adult-like symbols, and if they do, how they acquired them. In other words, even though these models are more advanced than psycholinguistic models in formally addressing the effectiveness of the proposed learning algorithms, it is not clear that they are solving the same bootstrapping problem than the one faced by infants. In addition, they typically lack a connection with empirical data on developmental trajectories.[3]

3.3 Developmental Artificial Intelligence

The idea of using computational models to shed light on language acquisition is as old as the field of cognitive science itself, and a complete review would be beyond the scope of this paper. We mention some of the landmarks in this field which we refer to as *Developmental AI*, separating three learning subproblems: syntax, lexicon, and speech. Computational models of syntax learning in infants can be roughly classified into two strands, one that learns from strings of words alone, and one that additionally uses a conceptual representation of the utterance meaning. The first strand is illustrated by Kelley (1967). The proposed computational model performed hypothesis testing and constructed more and more complex syntactic rules to account for the distribution of words in the input. The input itself was artificial (generated by a context free grammar) and part of speech tags (nouns, verbs, etc.) were provided as side information. Since then, manual tagging has been replaced by automatic tagging using a variety of approaches (see Christodoulopoulos, Goldwater, & Steedman, 2010 for a review), and artificial datasets have been replaced by naturalistic ones (see D'Ulizia, Ferri, & Grifoni, 2011, for a review). This strand views grammar induction as a problem of representing the input corpus with a grammar in the most compact fashion, using both a priori constraints on the shape and complexity of the grammars and a measure of fitness of the grammar to the data (see de Marcken, 1996 for a probabilistic view). The second strand can be traced back to Siklossy (1968), and makes the radically different hypothesis that language learning is essentially a translation problem: children are provided with a parallel corpus of speech in an unknown language, and a conceptual representation of the corresponding meaning. The Language Acquisition System (LAS) of Anderson (1975) is a good illustration of this approach. It learns context-free parsers when provided with pairs of representations of meaning (viewed as logical form trees) and sentences (viewed as a string of words, whose meaning are known). Since then, algorithms have been proposed to learn directly the meaning of words (e.g., cross-situational learning, see Siskind, 1996), context-free grammars have been replaced by more powerful ones (e.g. probabilistic Combinatorial Categorical Grammar), and sentence meaning has been replaced by sets of candidate meanings with noise (although still generated from linguistic annotations) (e.g., Kwiatkowski, Goldwater, Zettlemoyer, & Steedman, 2012). Note that all of these models take textual input, and therefore make the (incorrect) assumption that infants are able to represent their input in terms of an error-free segmented string of words.

The problem of word learning itself has been addressed using two main ideas. One main idea is to use distributional properties that distinguish within word and between word phoneme sequences (Harris, 1954; Elman, 1990; Chris-

[3] A particular difficulty of formal models which lack of a processing component is the observed discrepancies between the developmental trajectories in perception (e.g. early phonotactic learning in 8-month-olds) and production (slow phonotactic learning in one to 3-year-olds).

tiansen, Conway, & Curtin, 2005). A second idea, is to simultaneously build a lexicon and segment sentences into words (Olivier, 1968; de Marcken, 1996; Goldwater, 2007). These ideas are now frequently combined (Brent, 1996a; M. Johnson, 2008). In addition, segmentation models have been augmented by jointly learning the lexicon and morphological decomposition (M. Johnson, 2008; Botha & Blunsom, 2013), or tackling phonological variation through the use of a noisy channel model (Elsner, Goldwater, & Eisenstein, 2012). Note that all of these studies assume that speech is represented as an error-free string of adult-like phonemes, an assumption which cannot apply to early language learners. Finally, some studies have addressed language learning from raw speech. These have either concerned the discovery of phoneme-sized units, the discovery of words, or both. Several ideas have been proposed to discover phonemes from the speech signal (self organizing maps: Kohonen, 1988; clustering: Pons, Anguera, & Binefa, 2013; auto-encoders: Badino, Canevari, Fadiga, & Metta, 2014; HMMs: Siu, Gish, Chan, Belfield, & Lowe, 2013; etc.). Regarding words, D. K. Roy and Pentland (2002) proposed a model that learn both to segment continuous speech into words and map them to visual categories (through cross situational learning). This was one of the first models to work from a real speech corpus (parents interacting with their infants in a semi-directed fashion), although the model used the output of a supervised phoneme recognizer. The ACORNS project (Boves, Ten Bosch, & Moore, 2007) used real speech as input to discover candidate words (Ten Bosch & Cranen, 2007, see also Park & Glass, 2008; Muscariello, Gravier, & Bimbot, 2009, etc.), or to learn word-meaning associations (see a review in Räsänen, 2012). In sum, developmental AI represents the clearest attempt so far of addressing the full bootstrapping problem. Yet, although one can see a clear progression, from simple models and toy datasets, towards more integrative algorithms and more realistic datasets, there is no single proposition yet that handles the entire speech processing pipeline, i.e., from signal to semantics. Until this is done, it is not clear how the bootstrapping problem as faced by infants can be solved. In addition, the progression has been very discontinuous across studies, and in our view, hampered by the complete lack of cumulativity in algorithms, evaluation methods and corpora, overall making it impossible to compare the merits of the different ideas and register progress. Finally, even though most of these studies mention infants as a source of inspiration of the models, almost none of them try to account for developmental trajectories.

3.4 Summing up

Conceptual psycholinguistic models try to account for developmental trajectories but are not specified enough for demonstrating that they can solve the bootstrapping problem. Specific learning mechanisms address bootstrapping issues but only apply to toy or experimental data and cannot demonstrably scale up. This limitation calls for the need to develop *effective computational models* that work at scale. Both linguistic models and developmental AI attempt to effectively address the bootstrapping problem, but make unrealistic assumptions with respect to the input data (linguistic models take only symbolic input data, and most developmental AI models take either symbolic data or simplified inputs). As a result, these models address a different bootstrapping problem than the one faced by infants. This would call for the need to use *realistic data* as input for models. Both linguistic models and developmental AI models take as their gold standard description of the stable state in adults. This may be fine when the objective is to explain ultimate attainment (the bootstrapping problem), but does not enable to connect with learning trajectory data. This would call for a direct *human-machine comparison*, at all ages. Finally, a problem with much past computational modeling research in general is that even though they enabled quantitative predictions, the resources used were both specific to each study and not distributed freely, making it difficult to compare the different ideas and build on them. This calls for *open sourcing* these resources. Obviously, the reviewed approaches have limits but also address part of the puzzles. They need to be combined, and the proper way to achieve this combination is examined next.

4 Four requirements

Here, we examine in more details the four requirements outlined above, and discuss them in the following order: using realistic data, comparing humans and machines, constructing effective computational models, open-sourcing data, evaluation and models. We conclude by discussing *biological plausibility* as a possible additional requirement.

4.1 Using realistic data

One of the most serious limitations of past theoretical work is the tendency to focus either on a simplified learning situation, a small corpus, or both, thereby failing to address the language leaning problem in its full complexity. Of course, simplification is the hallmark of the scientific enterprise, but we claim that in the present case, simplifications often result in the learning problem itself being distorted beyond recognition. We therefore argue that to address the bootstrapping problem, one has to use realistic data as input. Formal learning theory provides us with many examples where idealizing assumptions about the learning situation (regarding the input to the learner or the set of target languages to be learned) have extreme consequences on what can be learned or not. For instance, if the environment presents only positive instances of grammatical sentences presented in any possible order, then even simple

Table 1
Four studies used to estimate infant's speech input

study	reference	mode of acquisition;age	population
H&R	Hart and Risley (1995)	observer, 1h every month; 12-36 months	urban high, mid & low SES, English
SALG	Shneidman, Arroyo, Levine, and Goldin-Meadow (2013)	observer, 1h every month; 12-36 months	urban high SES, English & rural low SES, Maya
W&F	Weisleder and Fernald (2013)	daylong recording; 19 months	low SES, Spanish
VdW	van de Weijer (2002)	daylong recording; 6-9 months	high SES, Dutch

Table 2
Estimates of yearly input, in total, and restricted to Child Directed Speech (CDS) , in number of hours and words (millions) per year in four studies (see the references in Table 1) as a function of sociolinguistic group (SES: Socio Economic Status). The numbers between brackets provide the range [min, max] of these numbers across families. ^{t} uses a wake time estimate of 9 hours per day. ^{w} uses a word duration estimate of 400ms. ^{c} uses SALG's estimate of %CDS for high SES. ^{d} uses W&F's estimate of %CDS for low SES. ^{m} uses H&R's MLU's estimates (according to SES).

	Yearly total				Yearly CDS			
	Hours		Words (M)		Hours		Words (M)	
				Urban, high SES				
H&R (N=13)^t	$1221^{w,c}$	[578,1987]	11.0^c	[5.20, 17.9]	786^w	[372, 1279]	7.07	[3.35, 11.5]
SALG (N=6)^t	$2023^{w,m}$	[1243, 2858]	18.2^m	[11.2, 25.7]	$1223^{w,m}$	[853, 1574]	11.0^m	[7.7, 14.2]
VdW (N=1)	931		9.28		140		1.39	
				Urban, low SES				
H&R (N=6)^t	$363^{w,d}$	[136, 558]	3.26^d	[1.22, 5.02]	225^w	[84, 346]	2.02	[0.76., 3.11]
W&F (N=29)^t	363^w	[52, 1049]	3.27	[0.46., 9.44]	225^w	[32, 650]	2.03	[0.29, 5.85]
				Rural, low SES				
SALG (N=6)^t	$503^{w,m}$	[365, 640]	4.53^m	[3.28, 5.76]	$234^{w,m}$	[132, 322]	2.10^m	[1.19, 2.90]

classes of grammars (e.g., finite state or context free grammars, Gold, 1967) are unlearnable. In contrast, if the environment presents sentences according to processes that can be recursively enumerated (an apparently innocuous requirement), then even the most complex classes of grammars (recursive grammars)[4] become learnable. This result extends to a probabilistic scenario where the input sentences are sampled according to a statistical distribution: constraints about the shape of the distribution radically changes the difficulty of the learning problem (see Angluin, 1988). In addition, the presence of *side information* can make a substantial difference: providing the syntactic trees along with the phonological form can turn an unlearnable problem into a learnable one (Sakakibara, 1992). The scale of the dataset can also have drastic effects, even when realistic data is used. This is illustrated by the history of automatic speech recognition systems. This field started to construct systems aimed at recognizing a small vocabulary for a single speaker (single digits) in the 50's, and nowadays handles multiple speakers with large vocabularies in spontaneous speech. By moving from small scale to big scale problems the field did not only use bigger models and more powerful machines, but had to build systems based on completely different principles (in order of appearance, formant based pattern matching, dynamic programming, statistical modeling, neural networks). Such heavy dependence on the scale and realism of the dataset is even more apparent with models of learning. For instance, dramatically different performances are found when word segmentation algorithms (which attempt to recover word boundaries from continuous speech) are fed with a phoneme transcription or when they are fed with raw speech signals (Jansen, Dupoux, et al., 2013; Ludusan, Versteegh, et al., 2014). Addressing the data scalability problem can be done according to two approaches. One approach is to work with idealized inputs generated by simple formal grammars or probabilistic models, which are made more complex incrementally to approximate real data. While this approach, pursued by formal learning theory, yield in-

[4] The problem of unrestricted presentations is that, for each learner, there always exists a 'nemesis', an evil environment that will trick the learner into converging on the wrong grammar (see Clark & Lappin, 2011 for a detailed explanation).

teresting learnability theorems, it has to face the fact that there is currently no known formal grammar that ultimately characterizes the class of human languages (e.g., Jäger & Rogers, 2012). Even if it were the case, the particular presentation of the target language and associated side information that result from caretaker's communicative and pedagogic intentions has not been formally characterized. This approach therefore runs the risk of locking researchers in a bubble universe where idealized learning problem can be shown to be tractable, but are unrelated to that faced by infants in the real world. The second approach, which we promote as "reverse engineering" takes a radical step: instead of relying on formal descriptions of possible inputs, it uses *actual, attested, raw data* to reconstruct infant's sensory input. Here, it is still possible to do idealizations, but they would be expressed in terms of what transformation is done to the raw data before being fed to the learning system (for instance: selecting frequency bands based on the capacity of the auditory system, performing auditory scene analysis, etc.). We discuss three important consequences of this proposed solution: qualitative, quantitative, and cross-linguistic. On the *qualitative* side, as the input is now defined in terms of the sensory experience of the learner, there is not necessarily a predefined or preformatted language 'channel'. The reason for this is that the linguistic signals emitted by the parents are typically mixed with a variety of non linguistic signals in a culture dependent way. In addition, the physical medium of linguistic signals also vary from culture to culture. In the audio channel for instance, speech sounds are heard by infants mixed with all manners of background noise, music, and non linguistic vocal sounds. Within vocal sounds, click noises are considered non linguistic in many languages, but some languages use them phonologically (Best, McRoberts, & Sithole, 1988). In the visual channel, some amount of linguistic/communicative signals (gestures, mouth movements) is present in all cultures (Fowler & Dekle, 1991; Goldin-Meadow, 2005), but it becomes the dominant language channel in deaf communities using sign language (Poizner, Klima, & Bellugi, 1987). However, sign language can be used as native language even in hearing children, provided they are raised in mixed hearing/deaf communities (Van Cleve, 2004). Cross-cultural variation makes it impossible to innately specify a fixed way of unmixing these signals or selecting a language channel. It is therefore part of the language learning problem to separate the linguistic signals from the non-linguistic background. Using realistic inputs, instead of idealized ones would also expose the learner to linguistic signals that are corrupted or partially masked by other linguistic or non-linguistic signals. This cannot be entirely be dealt with through low level processing, as it is know that auditory source separation interacts with speech recognition (e.g. Warren, 1970). Similarly, dysfluencies and speech errors at many levels (Fromkin, 1984), individual differences (e.g., Hillenbrand, Getty, Clark, & Wheeler, 1995) and sociol-linguistics variation (Labov, 1972) are all factors are intgral to the learning problem and cannot be considered independantly solved. Yet, realistic inputs may also bring about potential benefits through 'side information'. As an example, syntax learning could be helped through the detection of prosodic information present in the signal. Prosodic boundaries may not always be coincidental with syntactic boundaries, but they could provide to the learner useful side information for the purpose of syntax and lexical acquisition (e.g. Christophe, Millotte, Bernal, & Lidz, 2008; Ludusan, Gravier, & Dupoux, 2014). Similarly, semantic information in the form of visually perceived objects or scenes and afferent social signals may help lexical learning (D. K. Roy & Pentland, 2002) and help bootstrap syntactic learning (the semantic bootstrapping hypothesis, see Pinker, 1984).On the *quantitative* side, it is important that the *totality* of the input is being considered for the following reasons. First, it sets up boundary conditions for the learning algorithms. Algorithms that require more input than is generally available to infants can be ruled out. As an example, current distributional semantic models use between 3 and 100 billion words to learn vector representations for the meaning of words or short phrases based on adjacent words (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; *Word2vec Google Project Page*, 2013). This is between 30 and 1000 times more data than infants are typically exposed to during their first 4 years of life, in fact, more than most people get in a lifetime, and therefore not plausible as the sole mechanism for meaning learning. Vice versa, an algorithm that would require only 10% of what infants get would display superhuman capacities and therefore not be a good model either. Second, establishing variability in input, and in particular, the range of extreme variation would enable to quantify the *resilience* that human infants display, and which have to be matched by a successful computational model. See Tables 2 and 1 for an estimation of the amount of speech data available to infants in different cultures. On the *cross-linguistic* side, a successful model of the learner should not demonstrate learning for only one input dataset, but it should learn for any input dataset in any possible human language in any modality (see the equipotentiality criterion in Pinker, 1987). Since, as we argued above, the class of all possible language inputs it still not formally characterized, one could take the approach of *sampling* from a finite but ever expanding set of existing linguistic communities. An adequate sampling procedure would insure that, statistically speaking, a given computational model is (or is not) able to learn from any possible input. Practically speaking, it may be interesting to sample typologies and sociolinguistic groups in a stratified fashion to avoid over-fitting the learning model to a restricted set of learning situations.To sum up, using realistic input is the only way to make sure that modelers are addressing the right learning problem. This has significant consequences regarding the size of the dataset that has to be collected : complete sensory coverage over the first 3 or 4 years of life, for a representative sample of children over a representative sample of languages. Before such a dataset is available, of course, it is still interesting to use as proxy a variety of smaller or simplified datasets, provided that the sources of simplification are clearly stated and their implications for generalizability discussed.

4.2 Evaluating systems through human-machine comparison

For a modeling enterprise of any sort, it is important to specify a success criterion. A lingering limitation of past theoretical work is that too many distinct success criteria have been used. In fact, the diversity is so great that it is nearly impossible to compare the different propositions across research fields (and sometimes even within field), and to reach the same standards as cumulative science. For psycholinguistic conceptual frameworks, the primary success criterion is the ability to account for developmental trajectories. Because of the verbal nature of these frameworks, it can only be checked at an intuitive and qualitative level. For linguistic formal learning models, the main focus is the learnability puzzle and is usually defined in terms of *learnability in the limit* (Gold, 1967): A learner is said to learn a target grammar in the limit, if after *some* amount of time, his own grammar becomes equivalent to the target grammar. This standard formulation has been criticized as too lax (K. Johnson, 2004). Since there is no time limit on convergence, a learner that needs a million year's worth of data to converge would still be deemed successful. We know that most children converge on an adult grammar in a fixed number of years, which is bounded by puberty. Therefore, our learnability criterion should be stronger and require the system to converge on a grammar after the *same amount of input* that it takes for children to converge. In addition the standard criterion assumes that one can determine when two grammars are equivalent, which may not be tractable.[5] Finally, for the developmental AI models that we reviewed, system evaluation was not their strong selling point. Many provided only qualitative evaluations, but for those that did provide a numeric one, they were typically defined in relation to a so called *gold standard*, i.e. human annotations (like phoneme transcriptions, part of speech annotations, parse trees, etc). The success of the learning algorithm is then measured as a distance between the machine annotation and the gold one. Of course, these evaluations are only valid to the extent that the gold standard reflects the state of the human language competence. This is not necessarily the case for adult-machine comparisons, as linguists may disagree on some of the annotations, and certainly not the case for children-machine comparisons, as the infant's grammar is probably different from that of the linguistically-trained adult. We therefore claim that for the reverse engineering approach, none of these criteria, taken individually, are satisfactory. Prior advocates of the use of machine learning to model language acquisition have proposed a number of ways to combine these criteria. To quote a few, MacWhinney (1978) proposed 9 criteria, Berwick (1985), 9 criteria (different ones), Pinker (1987) 6 criteria, Yang (2002) 3 criteria, M. C. Frank, Goldwater, Griffiths, and Tenenbaum (2010) 2 criteria. These can be sorted into conditions about effective modeling (being able to generate a prediction), about the input (being as realistic as possible), about the end product of learning (being adult-like) and about the plausibility of the computational mechanisms. In our proposed reverse engineering approach, we would like to integrate within a single operational criterion, the *cognitive indistinguishability criterion*, the insights of the psycholinguistic theories with the quantitative evaluations of the formal and algorithmic models:

> A human and a machine are cognitively indistinguishable with respect to a given set of tests when they yield numerically similar results when ran on these tests.

The proposal, therefore, is that, a computational model of language learning is successful, when it yields a system that is cognitively indistinguishable from a human (adult or child) after having been fed with the same input data. Such a success criterion enables both to address the learnability puzzle and to account for developmental trajectories. Note, however, that cognitive indistinguishability is not an absolute criterion but depends on a set of tests. Constructing an agreed upon set of such tests (a *cognitive benchmark*) becomes therefore part of the reverse engineering project by integrating tests that linguists and psycholinguists agree upon as being relevant to characterizing grammatical competence in humans. This benchmark can of course be revised as new and more subtle experimental protocols for language competence are discovered and can set the human and machine apart. Here, we present three conditions that such tests must satisfy to achieve our scientific objectives: they should be *administrable* (to adults, children and computers alike), *valid* (measure the construct under study as opposed to something else), and *reliable* (with a good signal to noise ratio). The last two conditions are common in psychometrics and psychophysics (e.g., Gregory, 2004). Test validity refers to whether a test, both theoretically and empirically, is sensitive to the psychological construct (state or process) it is supposed to measure. For instance, in an influential paper,

[5] Two grammars are said to be weakly equivalent if they generate the same utterances. In the case of context free grammars, this is an undecidable problem. More generally, for many learning algorithms (e.g., neural networks), it is not even clear what has been learned, and therefore the criterion cannot be verified.

Turing (1950) proposed to test whether machines can 'think' using the so-called *imitation game*, where it had to persuade a human observer that it was a *female human* through an online keyboard conversation. The machine succeeds if it fools the observer as often as a human male participant would. This test is evidently not valid, as theoretically, 'thinking' is not a well defined psychological construct, but rather a polysemous folk psychology concept, and empirically, it is rather easy to fool human observers using rather simplistic text manipulation rules (see ELIZA, Weizenbaum, 1966). Fortunately, since the 50's, cognitive psychology has progressed tremendously and can offer a rich set of valid tests for the evaluation of language-related cognitive components (see Section 5.2). Test reliability refers to the signal to noise ratio of the measure. It can be evaluated by rerunning the same tests over the same or different participants for humans, or over different initial conditions for the machines. Typically, test reliability is not thought to be a real issue for machines, to the extent that many algorithms are deterministic or assumed to be quite stable. Yet, it is important to assess this reliability empirically, for instance, by running the same algorithm over different samples of a large corpus. As for humans, test reliability is a very important issue, and even more so, for children and infants. Evidently, we cannot ask that the match between humans and machines be larger that the match within population. Test administrability does not belong to standard psychometrics, but it is especially important in the case both of infants and machines. Human adults have metalinguistic abilities which allow the experimenter to explain to them how to perform a particular test, in simple words. Such a strategy is not directly applicable to human infants nor to machines. In infants, a testing apparatus has to be constructed, i.e., a rather artificial environment whereby everything is controlled so that the response to test stimuli arises naturally and is measured using spontaneous tendencies of the participants (preference methods, habituation methods, etc; see Hoff, 2012, for a review).[6] In machines, there is also an issue of administrability. Typically, learning algorithms are not constructed to run linguistic tests, but to learn based on their input. Therefore, they need to be supplemented with particular *task interfaces* for each of the proposed tests in order to extract a response that would be equivalent to the response generated by humans.[7] In both cases, administering the task has to be made so as not to compromise the test's validity. Biases or knowledge of the desired response has to be removed from the testing apparatus (for the infants) and from the interface (for the machine). To sum up, to evaluate computational models, the reverse engineering approach proposes to build a revisable benchmark of valid and reliable tests measuring the various components of the human language faculty, and that can be administered to humans of various ages and machines alike. Models will be compared on their ability to mimic the results of these tests.

4.3 Constructing effective computational models

As discussed in Section 3, past work in psycholinguistics, formal linguistics and many studies in developmental AI were not centered on the task of building effective systems which would work with real data. As a result, they could not provide a proof of principle that the bootstrappping problem can be solved nor be used to generate quantitative predictions. However, we know that contructing effective processing system which deal with realistic input is possible: the recent successes of machine learning in speech and language tasks demonstrate it. Specifically, Speech and Language Technologies (SLT) is the area of engineering research devoted to construct systems that perform complex functions like converting speech to text (Automatic Speech Recognition), or conducting a simple question/answer dialogue with large scale noisy data (Natural Language Processing). These are behind the rise of voice services on smartphones (Siri, Cortana, etc). The main design feature of SLT systems is that even though they contain components that are related to psycholinguistic and linguistic levels of representations (for instance *acoustic models* incorporate phonetic and phonological information, *language models* incorporate lexical and syntactic information, *discourse models*, semantic and pragmatic information, and so on), it is not assumed that any of these representations are unambiguous and errorless. On the contrary, the handling of ambiguities and errors is built in from the ground up, through a processing architecture where, multiple and/or partial interpretations are passed in parallel from one level to the next along with their probabilities or activation levels, enabling the errors and ambiguities to be resolved in a holistic and optimal fashion (for a statistical framework in speech processing, see Jelinek, 1997, for a review of natural language processing, see (J. H. Martin & Jurafsky, 2008)). The problem is that these engineering systems are notat all models of the learning processes in infants. Instead, they are directly constructed as full-blown, performing adult. Furthermore, the way they are constructed uses a substantial amount of 'supervision', i.e., direct intervention of experts regarding how the system should be tuned depending on the language and application at hand. Early systems were heavily engineered, with each subcomponent crafted and tuned by hand by a team of experts. Nowadays, only a general architecture is specified, and the model parameters (in very large numbers) are tuned automatically using numerical optimization techniques run on

[6] In animals, before tests can be run, an extensive period of training is often necessary, in order for the animal to comply with the protocol. Such procedures are not possible in human infants.

[7] A task interface can be viewed as a function which takes as input the internal states of the algorithm generated by the stimuli and delivers a binary or real valued response.

(very large) datasets of human annotated speech or text. For instance, a typical state-of-the-art speech recognition component is trained with hours of hand transcribed speech (10000 hours or more), with a large pronunciation dictionary, and a few billion words of text. A language understanding component is trained with a bank of sentences annotated with part of speech and parse trees. All of these expert language resources are unfortunately not available to infants learning their native language(s). In brief, engineering approaches propose scalable processing systems, but even though they are based on statistical learning, they do not use cognitively plausible learning mechanisms because they rely on data not available to infants. The idea of the reverse engineering approach is, therefore, to use the general design principles and methods of SLT for robust and scalable processing, perhaps even adapting some of their algorithms with the aim at constructing systems that learn like infants do. To acheive this goal, two issues need to be addressed: the constructed system can learn without expert labels (using realistic data), the constructed system can be tested at every stage of development (using human benchmarks). Technically, learning mechanisms that only use raw signals (or indirect human feedback or 'labels') are called *unsupervised* (or *weakly supervised*). This class of machine learning problems is unfortunately less well studied and understood than the supervised learning ones (classification, regression, etc). Learning without external labels is obviously more difficult than learning with labels. Humans labels provide simultaneously a *target representation* that enable to evaluate how well the machine reproduce human competence, and an *error function* (the difference between the human provided and machine computed labels) that is optimized using numerical methods in order to reach this objective. With unsupervised problems, everything changes: the labels can still used to evaluate the machines's performance, but no longer for optimizing it. Indeed, the machine only has access to the inputs, and has to learn its own (so-called latent) representations of the input. This can also be written as an optimization problem, but the error function cannot refer to the labels (typically, the system's objective is to *model* its input, for instance, to predict future inputs based on past ones). Therefore the problem is much more underdetermined, and it is not clear that the latent representations will be anywhere near to the human labels. As for testability, the models have to be constructed so as to be able to process data at any intermediate learning state, including, in the initial state, ie, before any data has been presented at all. This precludes algorithms that optimize over an entire input corpus to directly construct a stable state in one step (example are segmentation algorithms that construct a lexicon by optimizing over the entire corpus, Brent, 1999). This, in contrast, favors algorithms that posess some form of incrementality (e.g., learn sentence by sentence, Pearl, Goldwater, & Steyvers, 2010, or day by day, assuming learning takes place during sleep). Oc course, the models have to be adapted, so as to be testable and can run experiments using test interfaces as described in Section 4.2.

To sum up, the best available option for constructing a scalable computational model of language learning comes from SLT systems, which needs to be refactored to work without expert supervision (no linguistic labels) in a weakly or unsupervised fashion, and to provide a testable processing system at all stages of development.

4.4 Open sourcing data, benchmarks and models

As any scientific endeavor, the reverse engineering approach proposed adheres to standard in transparency of process and replicability. As was noted above, many of the earlier attempts to bring machine learning to bear to issues of language development were too disconnected to allow cumulative science to proceed. It is therefore central for this proposal to share language datasets, test benchmarks and reference systems in an open source format to enable comparison of different models and enable new players to try their own ideas. Open source benchmarks, datasets and models are very common in many areas of machine learning, prominently in vision (e.g. the Imagenet dataset[8], Deng et al., 2009). This is less the case in speech and language, as many speech resources are protected or proprietary, thereby slowing down progress. Yet, this is changing quickly as open source speech databases are being constructed (for instance, the Librispeech dataset[9], Panayotov, Chen, Povey, & Khudanpur, 2015, and the Kaldi speech tools[10], Povey et al., 2011).

4.5 A biological plausibility requirement?

Here we briefly discuss one issue which often comes up when computational systems are used as models of human processing: the issue of *biological plausibility*. By this, we mean that the hypothetical algorithm be compatible with what we know about the biological systems that underlie these computations in human infants/adults. While this constraint is perfectly reasonable, we argue that it is difficult to apply to the modeling of early language acquisition for the following reasons: First, the computational power of a human brain is currently unknown. Current supercomputers can simulate at a synapse level only a fraction of a brain and several orders of magnitude slower than real time (Kunkel et al., 2014). If this is so, all computational models run in 2016 are still massively underpowered compared to a child's brain. Second, a particular algorithm may appear to be too complex for the brain, but a different version performing the

[8] http://www.image-net.org

[9] http://openslr.org

[10] http://kaldi-asr.org

same function will not. For instance, some word segmentation algorithms require a procedure called Gibbs sampling, which, in theory, require an infinite number of time steps to converge. This would seem to discredit the algorithm alltogether. Yet, it turns out that a truncated version of this algorithm running in finite time works reasonably well. Similarly, algorithms that require a lot of time steps can be rewritten into algorthms that require less steps and more memory. This makes a priori claims of biological plausibility difficult to make. Still, biological plausibility can place some theoretical bounds on *system complexity at the initial state*. Indeed, the initial state is constructed on the basis of the human genome plus prenatal interactions with the environment. This allows to rule out, for instance, a 100% nativist acquisition model that would pre-compile a state-of-the-art language understanding systems for all of the existing 6000 or more languages on the planet, plus a mechanism for selecting the most probable one given the input.[11] Apart from this rather extreme case, biological plausibility may not affect much of the reverse engineering approach until more is known about the computational capacity of the brain. Yet, it is compatible with our approach, since as soon as diagnostic tests of language computation in the brain are available, they could be added to the cognitive benchmark, as defined in Section 4.2.

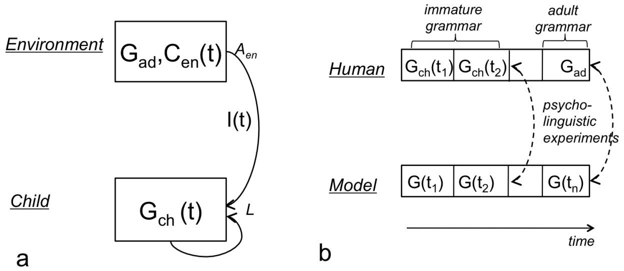

Figure 2. a. The (simplified) learning scenario: The Child's internal state is a grammar $G_{ch}(t)$ that can be updated through the learning function L based on input $I(t)$. The environment's internal state is a constant adult grammar G_{ad} and a variable context C_{en}, which produces the input to the child. b. Method to test the empirical adequacy of the model by comparing the outcome of psycholinguistic experiments with that of children and adults.

5 Feasibility and Challenges

We now turn to the feasibility of the reverse engineering approach as applied to early language acquisition. To do so, we first limit ourselves to the following simplifying scenario: the total input available to a particular child provides enough information to acquire the grammar of the language present in the environment. This may seem an innocuous assumption, but it essentially puts us in the open loop situation described in Figure 2), where the environment delivers a fixed curriculum of inputs (utterances and their sensory contexts) and the learner recovers the grammar that generated the utterances. In this situation, the output of the child is not modeled, and the environment does not modify its inputs according to her behavior or inferred internal states. We come back to this simplifying assumption in Section 7. Within this framework, we discuss how the four requirements can be met using current technology and the possible roadblocks that arise in the process of deploying this technology. As the fourth requirement is a methodological one which applies to the first three, we will not be discussing it separately. We will therefore review, in turn, the feasibility and challenges of data collection, testing and modeling.

5.1 Data Collection and Privacy

The requirement of using realistic data as input to the learner raises two issues, one technological and one ethical. At the technological level, it has become relatively easy to record virtually unlimited amounts of good quality audio and video data in children's environments. Perhaps the most ambitious data collection effort so far has been done within the Speechome project (D. Roy, 2009), where video and audio equipment was installed in each room of an apartment, recording 3 years' worth of data around one infant. Wearable recorders (see for instance the LENA system, Xu et al., 2008) enable recording the infant's sound environment for a full day at a time, even outside the home. These can be supplemented with position sensors to categorize activities (Sangwan, Hansen, Irvin, Crutchfield, & Greenwood, 2015), or Life logging wearable devices to capture images every 30 seconds in order to reconstruct the context of speech interactions (Casillas, 2016). Of course, part of the technological challenge is not only to record raw data, but also to reconstruct the infant's sensory experience, from a first person point of view. In this context, head-mounted cameras can be useful to estimate the infant's head (and therefore average gaze) direction (L. B. Smith, Yu, Yoshida, & Fausey, 2015). Recent progress in 3D reconstruction, especially when using multi-view and/or depth sensors make it possible to go further in sensory reconstruction (e.g., Mustafa, Kim, Guillemaut, & Hilton, 2016) although this has not yet been done with infant data. Finally, even raw sensory data is difficult to use if it not supplemented with reliable linguistic/high level annotations. For instance, a large part of the

[11] The reason such system would not be biologically realizable is that the parameters of a state-of-the-art phoneme recognition system for a single of these languages already require 10 times more memory storage than what is available in the fraction of the genome that differentiate humans from apes. A DNN-based phone recognizer has typically more than 200M parameters, which barring ways to compress the information, takes 400Mbytes. The human-specific genome is 5% of 3.2Gbase, which boils down to only 40Mbytes.

Speechome corpus's audio track has been transcribed using semi-automatized means, enabling the search for linguistic characteristics of both the input to the child and its output (B. C. Roy, Frank, DeCamp, Miller, & Roy, 2015). Continuous progress in machine learning (speech recognition: Amodei et al., 2015b; object recognition: Girshick, Donahue, Darrell, & Malik, 2016; action recognition: Rahmani, Mian, & Shah, 2016; emotion recognition: Kahou et al., 2015) will enable to lower the burden on high-level annotation of large amounts of data. The technological aspect of massive data collection, however, appears relatively simple when compared with the ethical challenges raised by the need to make this data accessible to the research community. There is a tension between the requirement of sharability and open scientific data (see Section 4.4), and the need of protecting individual privacy when it comes to personal and sensitive data. Up to now, the response of the scientific community has been dichotomous: either make everything public (as in the open access repositories like CHILDES, MacWhinney, 2000), or completely close off the corpora to anybody outside the institution that has recorded the data (as in the Riken corpus, Mazuka, Igarashi, & Nishikawa, 2006, or the Speechome corpus D. Roy, 2009). The first strategy sacrifices privacy and is impossible to scale up to dense recordings. The second strategy puts such an obstacle to the scientific use of the corpora that it almost defeats the purpose of conducting the recording in the first place. A number of alternative strategies are being considered by the research community. The Homebank repository contains raw and transcribed audio, with a restricted case by case access to researchers (VanDam et al., 2016). Databrary has a similar system for video recordings (https://nyu.databrary.org). Progress in cryptographic techniques would make it possible to envision preserving privacy while enabling more open exploitation of the data. For instance, the raw data could be locked on secure servers, thereby remaining accessible and revokable by the infants' families. Researchers' access would be restricted to anonymized meta-data or aggregate results extracted by automatic annotation algorithms. *Differential privacy* techniques enable outside participants to make queries on databases while providing a level of guarantee on the amount of private information that can be extracted (Dwork, 2006). The specifics of such a new type of linguistic data repository would have to be worked out before dense speech and video home recordings can become a mainstream tool for infant research. In brief, massive data collection is technically feasible, but it's exploitation in an open source format requires specific developments in privacy-preserving storage and computing infrastructures.

5.2 Cognitive Benchmarking and Experimental Reliability

Our second requirement, the construction of a cognitive benchmark for language processing, can be considered a done thing in the case of the human adult. The linguistic and psycholinguistic communities have indeed constructed relatively easy-to-administer, valid and reliable tests of the main components of linguistic competence in perception/comprehension (see Table 3). These tests are easy to administer because they are conceptually simple and can be administered to naive participants; most of them are of two kinds: goodness judgments (say whether a sequence of sound, a sentence, or a piece of discourse, is 'acceptable', or 'weird') and matching judgments (say whether two words mean the same thing or whether an utterance is true of a given situation, which can be described in language, picture or other means). The validity of linguistic tests often stems from the fact that they are used within a *minimal set design*. Such design selects examples where only one linguistic construct is manipulated while every other variable is kept constant (for instance: 'the dog eats the cat' and 'the eats dog the cat' only differ in word order). Regarding test reliability, as it turns out, many linguistic tests are quite reliable, as 97% of the results in a textbook of linguistics are replicable using on-line experiments (Sprouse, Schütze, & Almeida, 2013)[12]. Given the simplicity of these tasks, it is relatively straightforward to apply them to machines. Indeed, matching judgments between stimulus A and stimulus B can be derived by extracting from the machine the representations triggered by stimulus A and B, and compute a *similarity score* between these two representations. Goodness judgments are perhaps more tricky; they can easily be done by generative algorithms that assign a *probability score*, a *reconstruction error*, or a *prediction error* to individual stimuli. As seen in Table 3, some of these tests are already being used quite standardly in the evaluation of unsupervised learning systems, in particular, in the evaluation of phonetic and semantic levels) while for others they are less widespread.[13] Of course, in order to evaluate the ability of models to account for developmental trajectories (second puzzle) we must also compare machines with children. This is where the difficult challenge lies. The younger the child, the more difficult it is to construct reliable tests. The replicability crisis (see Ioannidis, 2012; Open Science Collaboration, 2015) has barely hit developmental

[12] This is a simplification of the situation: even in simple psychophysical tasks, humans can be affected by many other factors like attention, fatigue, learning or habituation to stimuli or regularities in stimulus presentations, etc. Methods try to minimize but never totally suceed in neutralizing these effects.

[13] Regarding the evaluation of word discovery systems, see the proposition by Ludusan, Versteegh, et al. (2014) but see Pearl and Phillips (2016) for a counter proposal and a discussion in (Dupoux, 2016).

Table 3
Example of tasks that could be used for a Cognitive Benchmark.

Task description in human adults	Linguistic level	Equivalent task in children	Equivalent task in machines
Well-formedness judgement *does utterance S sound good?*	phonetic, prosody, phonology, morphology, syntax	preferential looking (9-month-olds: Jusczyk, 1997), acceptability judgment (2-year-olds: de Villiers and de Villiers, 1972; Gleitman, Gleitman, and Shipley, 1972)	reconstruction error (Allen & Seidenberg, 1999), probability (Hayes & Wilson, 2008), mean or min log probability (Clark, Giorgolo, & Lappin, 2013)
Same-Different judgment *is X the same sound / word / meaning as Y?*	phonetic, phonology, semantics	habituation / deshabituation (newborns, 4-month-olds: Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Bertoncini, Bijeljac-Babic, Blumstein, & Mehler, 1987), oddball (3-month-olds: Dehaene-Lambertz, Dehaene, et al., 1994)	AX/ABX discrimination (Carlin, Thomas, Jansen, & Hermansky, 2011; Schatz et al., 2013), cosine similarity (Landauer & Dumais, 1997)
Part-Whole judgment *is word X part of sentence S?*	phonology, morphology	Word spotting (8-month-olds: Jusczyk, Houston, & Newsome, 1999)	spoken web search (Fiscus, Ajot, Garofolo, & Doddington, 2007)
Reference judgment *does word X (in sent S) refer to meaning M?*	semantics, pragmatics	intermodal preferential looking (16-month-olds: Golinkoff, Hirsh-Pasek, Cauley, & Gordon, 1987), picture-word matching (11-month-olds: Thomas, Campos, Shucard, Ramsay, & Shucard, 1981)	picture/video captioning (e.g., Devlin, Gupta, Girshick, Mitchell, & Zitnick, 2015), Winograd's schemas (Levesque, Davis, & Morgenstern, 2011)
Truth/Entailment judgment *is sent S true (in context C)?*	semantics	Truth Judgment Task (3-year-olds: Abrams, Chiarello, Cress, Green, & Ellett, 1978; Lidz & Musolino, 2002)	visual question answering (Antol et al., 2015)
Felicity judgement *would people say S to mean M (in context C)*?	pragmatics	Ternary reward task (5-year-olds: Katsos & Bishop, 2011), Felicity judgment task (5 years olds: Foppolo, Guasti, & Chierchia, 2012).	?

psychology yet because there are so few replications in the first place (although, see the Many Babies project, M. Frank, 2015). Addressing this challenge would require improving substantially the reliability of the experimental techniques. Existing meta-analyses highlight large differences in effect sizes across experimental methods (community-augmented meta-analyses: Tsuji, Bergmann, & Cristia, 2014, metalab: http://metalab.stanford.edu/), which point to ways to improve the methods. If the method's signal-to-noise reach a plateau, there is the possibility to increase the number of participants through collaborative testing, as in genome-wide association studies, where low power requires a consortium to run very large number of participants (e.g., around 200,000 participants in Ehret, Munroe, Rice, & al., 2011) or increase the number of data points per child (perhaps using home-based experiments: L. Shultz, 2014, https://lookit.mit.edu/, or V. Izard, 2016, https://www.mybabylab.fr). In brief, while a cognitive benchmark can be established, some of the fine grained predictions of reverse engineering models in infants will require progress in developmental experimental methods.

5.3 Unsupervised learning of speech and language understanding

The third requirement of constructing effective computational models of the learner faces one major challenge: the feasibility of unsupervised or weakly supervised learning, i.e. to learn latent linguistic representations instead of being force-feed these representations through expert annotations. Two main, non exclusive, ideas are being explored to address this challenge. One idea could be referred to under the generic name of *prior information*. It is the idea that one can replace some of the missing labels (expert information) by innate knowledge about the structure of the problem. With strong prior knowledge, some logically impossible induction problems become solvable.[14] The reasoning here is that evolution might have given the learning system strong prior knowledge about some universal regularities of language, such that only few data points are necessary to learn the rel-

[14] One good illustration is the following: can you tell the colors of 1000 balls in an urn by just selecting one ball? The task is impossible without any prior knowledge about the distribution of colors in the urn, but very easy if you know that all the balls have the same color.

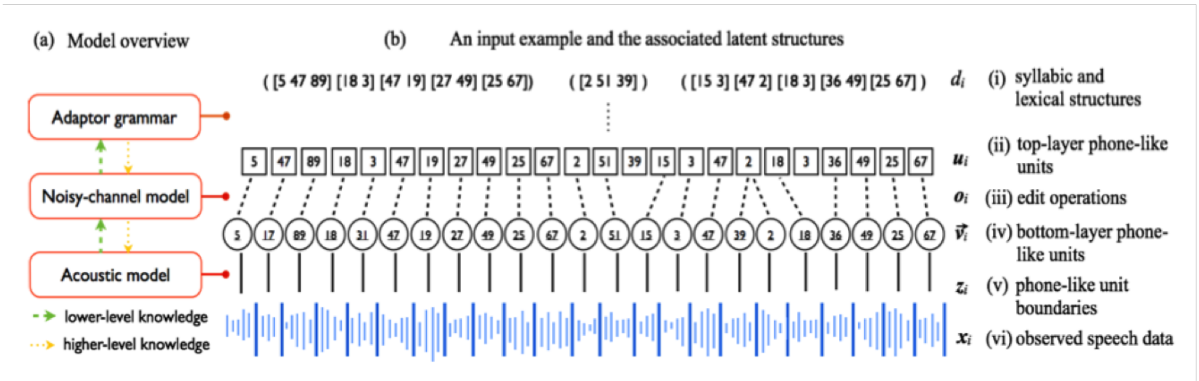

Figure 3. Outline of a generative architecture learning jointly words and phonemes from raw speech (from Lee, O'Donnell & Glass, 2015).

evant system. Such ideas have been proposed in the acquisition of syntax under the name of principles and parameters. Under this theory, a single sentence (called a trigger) is sufficient to decide on one parameter (Gibson & Wexler, 1994; Sakas & Fodor, 2012). An illustration of such a system for learning phonemes and words from raw speech uses a very specific generative architecture to guide the learning process (Lee & Glass, 2012; Lee, O'Donnell, & Glass, 2015, see Figure 3). The second idea is that of *soft constraints* coming from a large interconnected system. Instead of trying to learn each subcomponent of language in isolation, the idea is to integrate these subsystems in a general language processing architecture, and let the subcomponent constrain each other. Because each subcomponent is solving a different optimization problem, they are providing the other subsystems their own view of what has to be learned. For instance, in the domain of phonetic learning it has been shown that even an imperfect, automatically discovered lexicon can help improving on subword representations using allophonic representations (A. Martin, Peperkamp, & Dupoux, 2013; Fourtassi & Dupoux, 2014) or the raw speech signal (Jansen, Thomas, & Hermansky, 2013; Thiollière, Dunbar, Synnaeve, Versteegh, & Dupoux, 2015, see Figure 4). This idea has been discussed under different guises (multitask learning: Caruana, 1997; multi-cue integration: Christiansen et al., 2005), but is perhaps best expressed under the notion of learning *synergies* (M. Johnson, 2008). Synergies correspond to the fact that jointly learning two aspects of language is easier than learning either one alone.[15] They have been documented among others, between phonemes and words inventories (Feldman, Myers, White, Griffiths, & Morgan, 2011), syllables and words segmentation (M. Johnson, 2008), referential intentions and word meanings (M. C. Frank, Goodman, & Tenenbaum, 2009). Note that the envisioned solutions address squarely the puzzles mentioned in the introduction: learning takes place without supervisory signals (the unsupervised/weakly supervised setting), all levels are learned simultaneously (joint modeling), and language learning is resilient.

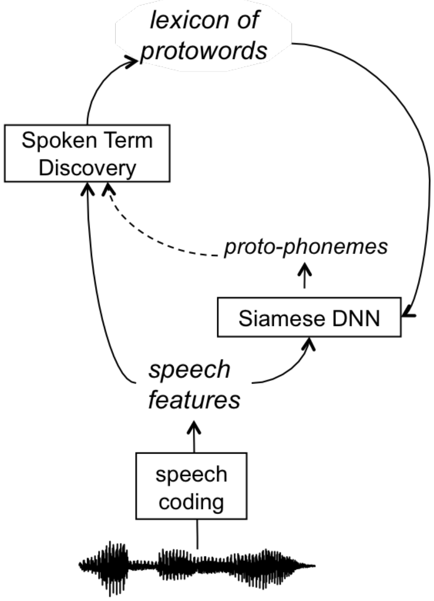

Figure 4. Architecture illustrating a top-down synergy between learning phonemes and words. Auditory spectrograms (speech features) are computed from the raw speech signal. Then, protowords are extracted using Spoken Term Discovery; these words are then used to learn a more invariant speech representation using discriminative learning in a siamese Deep Neural Network architecture (from Thiolliere et al., 2015).

This last point can be viewed as a correlate of the soft constraint idea: a (reasonable) limitation in some input can be compensated for by strong priors and/or information coming from another linguistic or non linguistic level. In brief, even though unsupervised/weakly supervised learning is difficult,

[15] Interestingly, the idea of synergies turn the bootstrapping problem on its head: instead of being a liability, the codependancies between linguistic components become an asset.

there is a growing interest within machine learning for the study of such algorithms. This opens up a window of opportunity for collaborations between the cognitive science and machine learning communities.

6 Preliminary results

Achieving the goals of the reverse engineering approach is a long-term project, requiring us to overcome the challenges listed in the preceding section. This means that we will have to contend with partial realizations for many years. Yet, even partial realizations would provide useful benefits in the area of cognitive and linguistic theories, corpus studies, experimental studies and machine learning. We illustrate these benefits next, through a selection of examples in 5 research areas.

6.1 Challenging intuitions and psychological theories

As discussed above, cognitive theories of the language learner come under the shape of verbally expressed *conceptual frameworks*, in which inferences and predictions are left to the interpretations and intuitions of the reader. The reverse engineering approach shows that in some cases, effective implementations of intuitive ideas yield counterintuitive results. As for specific learning mechanisms that have been modelled and tested with toy data, implementing and testing them at scale can be useful to assess the effectiveness of these mechanisms and their relative strength in real life situations. We briefly illustrate this with three examples. The first example is the learning of phonetic categories by infant through 'distributional learning' which can be viewed as a mechanism of unsupervised clustering. Even though this mechanism was validated in infants (Maye et al., 2002), and several implemented algorithms were tested (Vallabha et al., 2007; McMurray et al., 2009, among others), it appears that none of these tests were run on real continuous speech datasets. Most papers used either toy data (points in formant space generated from a Gaussian distribution), or worked from measurements made on manually segmented speech. When tested on continuous speech, clustering algorithms yield a very different result. For instance, Varadarajan, Khudanpur, and Dupoux (2008) have shown that a clustering algorithm based on Hidden Markov Models and Gaussian mixtures does not converge on phonetic segments, but rather, on much shorter (30 ms), highly context-sensitive acoustic events. To find phoneme-sized units would seem to require a different algorithm with strong priors on the temporal structure of phonemes (Lee & Glass, 2012). This example reveals that contrary to the hypothesis in Maye et al. (2002), finding phonetic units is not only a problem of constructing categories (clustering), it is also a problem of segmenting continuous speech. Furthermore, the two problems are not independent and have therefore to be addressed jointly by the learning algorithms. This, in turn, would yield specific predictions to be tested in infants. The second example is word segmentation using transition probabilities. Even though a lot of work has been devoted to study the importance of transition probabilities as a possibly cue to signal word boundaries in infants (Romberg & Saffran, 2010, for a review), it turns out that this cue alone yields disappointingly poor segmentation performance in a real corpus. In contrast, algorithms based on totally different principles which directly learn a lexicon, and obtain a segmentation as a by-product fare a lot better (Cristia et al. in preparation). Unfortunately, even though such lexical-based algorithms could potentially be much more useful for language acquisition, they have been little studied empirically in infants. The third example relates to the popular hypothesis that the meaning of words is acquired by infants through the coocurrence patterns of verbal material with contextual cues in other modalities (for instance, the presence of a dog when hearing the word 'dog'). Yet, Fourtassi and Dupoux (2014) has shown that it is possible to derive an approximate representation of the meaning of words without any cross modal information, through coocurrence patterns within the verbal material itself. Counterintuitively, such approximate meaning representation can provide useful top-down feedback on how to cluster phonetic information into phonemes. This shows that large scale computational models may suggest new types of Ãă priori implausible but effective mechanisms.

6.2 Grounding formal linguistic theories

Every linguistic theory relies on a core list of representations and symbols that are supposed universal. For instance, Optimality Theory relies on a list of universal phonetic features and constraints. The same goes with syntactic and semantic theories (part of speech, types of grammatical relations or computations, quantifiers, etc.). Where do these symbols come from and how they are grounded in the signal remains unspecified. Reverse Engineering offers the possibility to give an account of the developmental emergence of these elements. For instance, in the domain of phonology, Dunbar, Synnaeve, and Dupoux (2015) has proposed that phonological features could emerge from a joint auditory and articulation space. In the domain of lexical semantics, distributional accounts have emerged which ground the meaning of words into the unsupervised learning of patterns of concurrence (Landauer & Dumais, 1997). These patterns correlate well with judgments of semantic proximity (although see Linzen, Dupoux, & Spector, 2016; Gladkova, Drozd, Center, & Matsuoka, 2016). Similarly, in the domain of syntax, systems of automatically derived part of speech tags through unsupervised distributional learning finding work as well, and on some occasion better than those tags provided by experts (e.g., Prins & Van Noord, 2001). The potential consequences of these results for foundational issues in formal linguistic theories remain to be explored.

6.3 Characterizing the input

Corpus studies characterize the input to the child in terms of various measures of linguistic complexity (mean length of utterance, lexical diversity, etc). What reverse engineering can offer is a new set of tools to quantify linguistic complexity *with respect to its effect on the language learner*. We briefly give two examples, one cross linguistic, one that regards the so-called hyperspeech hypothesis. Regarding cross-linguistic variation, languages differ in great extent in the complexity of their surface features. How much do these variations matter to the learner? This can be explored by systematically running language learning algorithms through corpora of different languages. For instance, Fourtassi, Boerschinger, Johnson, and Dupoux (2013) have replicated in a controlled fashion the often noted finding that word segmentation models make a lot more errors in some languages than others (e.g., Japanese versus English). They showed that the difference in performance was not specific to the algorithm that they used, but was related an intrinsic difference in segmentation ambiguity between the two languages, which is itself based on their differing syllabic structure. Conducting similar studies cross-linguistically would help to derive a new learnability-based linguistic typology, which could then be related to potential cross-linguistic differences in learning trajectories in infants. As for the hyperspeech hypothesis, it was proposed that parents adapt their pattern of speech to infants in order to facilitate learning (see Fernald, 2000 for a discussion). Consistent with this, Kuhl (1997) observed that parents tend to increase the separation between point vowels in child directed speech, possibly making them more distinctive. Yet, Ludusan, Seidl, Dupoux, and Cristia (2015) ran a word discovery algorithm on raw speech and failed to find any difference in word learning between child and adult directed speech; if anything, the former was slightly more difficult. This paradoxical result can be explained by the fact that parents tend to increase phonetic variability when addressing their infants, which results in a net decrease in *category discriminability* (A. Martin et al., 2015; see also McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013). Scalable computational models are therefore useful to assess the net functional role of otherwise disparate linguistic and phonetic effects. More topics could be explored following the same approach. For instance, some studies have proposed that parents provide informative feedback even on preverbal vocalization (Gros-Louis, West, Goldstein, & King, 2006; Plummer, 2012; Warlaumont, Richards, Gilkerson, & Oller, 2014). A modeling approach would help to determine whether such behavior truly helps language learning in a naturalistic environment. The same goes for other forms of weak parental supervision like referential pointing, joint attention, etc.

6.4 Errors as predictions

Before a full model of the learner is available, even a partial model would enable to provide useful predictions. For instance, even the best unsupervised segmentation mechanisms fed with errorless phonemic transcriptions make systematic errors: under-segmentations for frequent pairs of words (like "readit" instead of "read"+"it") or over-segmentations (which would be "butterfly" being segmented into "butter"+"fly") (see Peters, 1983). Instead of viewing these errors are inadequacies of the models, one could view them as reflecting areas of the target language that are intrinsically difficult to segment in the absence of other information (syntactic, semantic, etc). Therefore, it is reasonable to expect that infants would make the same errors, at least at an age where the assumptions of the models are met (after having stabilized their phonetic representations, but before much semantic/syntactic learning). These 'errors' could then be presented into infants and tested for recognition. If infants are making the same errors as the proposed mechanism, this would count as evidence in favor of this mechanism. Ngon et al. (2013) tested the prediction of a very simple model of word segmentation (an ngram model) run on a CHILDES corpus. Eleven month olds preferred to listen to some frequent mis-segmentations of the model, and did not distinguish them from real words of the same frequency. The logic could be extended by running different models on the same data, and generating *diagnostic patterns* that distinguish between the competing models, allowing to separate them empirically. In principle, such diagnostic patterns could be generated, cross-linguistically, within a language, or even within a given individual infant (to the extent that the input data can be collected individually). The diagnostic pattern technique opens up therefore a whole arena for comparing implemented models and theories.

6.5 Collaborations with the Machine Learning community

As the reverse engineering approach develops cognitive benchmarks, this can provide new playgrounds (problem sets) for developing architectures and algorithms that can work with little or no supervision, with a moderate amount of data. Infants provide a proof of principle that such systems can be constructed. One example of this is the zero resource speech challenge (Versteegh et al., 2015) which explores the unsupervised discovery of sublexical and lexical linguistic units from raw speech. Such a challenge, set up with open-source datasets and baselines (see www.zerospeech.com) attracted considerable interest in the community of speech technology (Versteegh, Anguera, Jansen, & Dupoux, 2016). Such so-called zero-resource algorithms (Glass, 2012; Jansen, Dupoux, et al., 2013) are not only interesting models of infant early phonetic and lexical acquisition, they can also provide technical solutions for the

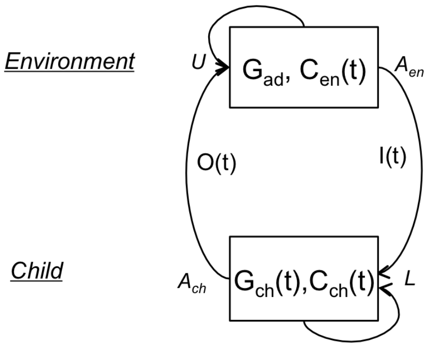

Figure 5. The learning situation in the interactive scenario, viewed as two coupled dynamic systems: the Child and the Environment.

construction of speech services in languages with scarce linguistic resources, or in languages with no or unreliable orthography.

7 Can reverse engineering address the fully interactive learning scenario?

The feasibility section (Section 5) endorsed a set of simplifying assumptions encapsulated in Figure 2a. This scenario does not take into consideration the child's output, nor the possible feedback loops from the parents based on this output. Many researchers would see this as a major, if not fatal, limitation of the approach. In real learning situations, infants are also agents, and the environment reacts to their outputs creating feedback loops (Bruner, 1975, 1983; MacWhinney, 1987; Snow, 1972; Tamis-LeMonda & Rodriguez, 2008). The most general description of the learning situation is therefore as in Figure 5. Here, the child is able to generate observable actions (some linguistic, some not) that will modify the internal state of the environment (through the monitoring function). The environment is able to generate the input to the child as a function of his internal state. In this most general form, the learning situation consists therefore in two *coupled dynamic systems*.[16] Could such a complex situation be addressed within the reverse engineering approach? We would like to answer with a cautious yes, to the extent that it is possible to adhere to the same four requirements, i.e., realistic data (as opposed to simplified ones), explicit criteria of success (based on cognitive indistinguishability), scalable modeling (as opposed to verbal theories or toy models) and sharable resources. While none of these requirements seem out of reach, we would like to pinpoint some of the difficulties, which are the source of our caution. Regarding the data, the interactive scenario would require accessing the full (linguistic and non linguistic) output of the infant, not only her input. While this is not intrinsically harder to collect than the input, and is already been done in many corpora for older children, the issue of what to categorize as linguistic and non linguistic output and how to annotate it is not completely trivial. Regarding computational modeling, instead of focusing on only one component (the learner) of one agent (the child), in the full interactive framework, one has to model two agents (the child and the adult) for a total of four components (the learner, the infant generator, the caretaker monitor, and the caretaker generator). Furthermore, the internal states of each agent has to be split into linguistic states (grammars) and non-linguistic (cognitive) states to represent the communicative aspects of the interaction (e.g., communicative intent, emotional/reinforcement signals). This, in turn, causes the split of each processing component into linguistic and cognitive subcomponents. Although this is clearly a difficult endeavor, many of the individual ingredients needed for constructing such a system are already available in the following research areas. First, within speech technology, there are available components to build a language generator, as well as the perception and comprehension components in the adult caretaker. Second, within linguistics, psycholinguistics and neuroscience, there are interesting theoretical models of the learning of speech production and articulation in young children (Tomasello, 2003; W. Johnson & Reimers, 2010; Guenther & Vladusich, 2012). Third, within machine learning, great progress has been made recently on reinforcement learning, a powerful class of learning algorithms which assume that besides raw sensory data, the environment only provides sporadic positive or negative feedback (Sutton & Barto, 1998). This could be adapted to model the effect of the feedback loops on the learning components of the caretaker and the infant. Fourth, developmental robotics studies have developed the notion of intrinsic motivation, where the agent actively seek new information by being reinforced by its own learning rate (Oudeyer, Kaplan, & Hafner, 2007). This notion could be used to model the dynamics of learning in the child, and the adaptive effects of the caretaker-child feedback loops. The most difficult part of this enterprise would perhaps concern the evaluation of the models. Indeed, each of these new components and subcomponents would have to be evaluated on their own in the same spirit as before, i.e., by running them on scalable data and testing them using human-validated tasks. For instance, the child language generator should be tested by comparing its output to age appropriate children's outputs, which requires the development of appropriate metrics (sentence length, complexity, etc) or human judgments. The cognitive subcomponents would have to be tested against experiments studying children and adults in experimentally controlled interactive loops (e.g., N. A. Smith & Trainor, 2008; Goldstein, 2008). In addition, because a complex system is more than the sum of its parts, individual component validation would not sufficient, and the entire system would

[16] We thank Thomas Schatz, personal communication, for proposing this general formulation.

have to be evaluated.[17] Fully specifying the methodological requirements for the reverse engineering of the interactive scenario would be a project of its own. It is not clear at present how much of the complications introduced by this scenario are necessary, at least to understand the first steps of language bootstrapping. To the extent that there are cultures where the direct input to the child is severely limited and/or the interactive character of that input circumscribed, it would seem that a fair amount of bootstrap can take place outside of interactive feedback loops. This is of course entirely an empirical issue, one that the reverse engineering approach should help to clarify.

8 Conclusion

During their first years of life, infants learn a vast array of cognitive competences at an amazing speed; studying this development is a major scientific challenge for cognitive science in that it requires the cooperation of a wide variety of approaches and methods. Here, we proposed to add to the existing arsenal of experimental and theoretical methods the reverse engineering approach, which consists in building an effective system that mimics infant's achievements. The idea of constructing an effective system that mimics an object in order to gain more knowledge about that object is of course a very general one, which can be applied beyond language (for instance, in the modeling of the acquisition of naive physics or naive psychology) and even beyond development. Related work exist in the area of computational neuroscience, which attempts to use machine learning architectures (deep learning) and bring it to bear to the analysis of neural representations for visual inputs (Cadieu et al., 2014; Isik, Tacchetti, & Poggio, 2016; Leibo, Liao, Anselmi, & Poggio, 2015). The computational rationality framework uses bayesian modeling to bring together the field of Artificial Intelligence and studies of human abilities like reasoning or decision making (Gershman, Horvitz, & Tenenbaum, 2015). Returning to language acquisition, we have defined four methodological requirements for this combined approach to work: using realistic data as input (which implies setting up sharable and privately safe repositories of dense reconstructions of the sensory experience of many infants), constructing a computational system at scale (which implies 'de-supervising' machine learning systems and turning them into models of infant learning), assessing success by running tests derived from linguistics on both humans and machines (which implies setting up cumulative benchmarks of cognitive and linguistic tests) and sharing all of these resources. We've argued that even before the challenges are all met, such an approach can help to understand how language bootstrap can take place in a resilient fashion, and can provide an effective way to derive quantitative predictions that are of interest both practically and theoretically. The reverse engineering approach we propose does *not* endorse a particular model, theory or view of language acquisition. For instance, it does *not* take a position on the rationalist versus empiricist debate (e.g., Chomsky, 1965, vs. Harman, 1967). Our proposal is more of a methodological one: it specifies what needs to be done such that the machine learning tools can be used to address scientific questions that are relevant for such a debate. It strives at constructing at least one effective model that can learn language. Any such model will both have an initial architecture (nature), and feed on real data (nurture). It is only through the comparison of several such models that it will be possible to assess the *minimal* amount of information that the initial architecture has to have, in order to perform well. Such a comparison would give a quantitative estimate of the number of bits required in the genome to construct this architecture, and therefore the relative weight of these two sources of information. In other words, our roadmap does not start off with a given position on the rationalist/empiricist debate, rather, a position in this debate will be an outcome of this enterprise.

Acknowledgments

This paper benefited immensely from comments from and discussion with Alex Cristia, Paul Smolensky, Gabriel Synnaeve and Thomas Schatz, and with all the members of the 'Bootphon' team and Laboratoire de Sciences Cognitives et Psycholinguistique. This work was supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the Ecole de Neurosciences de Paris, and the Region Ile de France (DIM cerveau et pensée).

References

Abrams, K., Chiarello, C., Cress, K., Green, S., & Ellett, N. (1978). Recent advances in the psychology of language. In R. Campbell & P. Smith (Eds.), (Vol. 4a, chap. The relation between mother-to-child speech and word-order comprehension strategies in children). New York: Plenum Press.

Allen, J., & Seidenberg, M. S. (1999). The emergence of grammaticality in connectionist networks. In B. MacWhinney (Ed.), *Emergentist approaches to language: proceedings of the 28th carnegie symposium on cognition* (p. 115-151). Lawrence Earlbaum Associates.

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., ... others (2015a). Deep speech 2: End-to-end speech recognition in english and mandarin. *arXiv preprint arXiv:1512.02595.*

[17] For instance, a combined learner/caretaker system should be able to converge on a similar grammar as a learner ran on realistic data. In addition, their interactions should not differ in 'naturalness' compared to what can be recorded in natural situations, see (Bornstein & Tamis-LeMonda, 2010).

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., … others (2015b). Deep speech 2: End-to-end speech recognition in english and mandarin. *arXiv preprint arXiv:1512.02595*.

Anderson, J. R. (1975). Computer simulation of a language acquisition system: A first report. In R. Solso (Ed.), *Information processing and cognition.* Hillsdale, N.J.: Lawrence Erlbaum.

Angluin, D. (1988). *Identifying Languages from Stochastic Examples* [Technical Report 614. New Haven, CT: Yale University].

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the ieee international conference on computer vision* (pp. 2425–2433).

Badino, L., Canevari, C., Fadiga, L., & Metta, G. (2014). An Auto-encoder based approach to unsupervised learning of subword units. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Bates, E., & MacWhinney, B. (1987). Competition, Variation and Language learning. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition* (pp. 157–193). Hillsdale, N.J.: Lawrence Erlbaum.

Bergelson, E., & Swingley, D. (2012, February). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253–3258.

Bertoncini, J., Bijeljac-Babic, R., Blumstein, S. E., & Mehler, J. (1987). Discrimination in neonates of very short cvs. *The Journal of the Acoustical Society of America*, *82*(1), 31–37.

Berwick, R. (1985). *The acquisition of syntactic knowledge*. MIT Press.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human perception and performance*, *14*(3), 345.

Bornstein, M. H., & Tamis-LeMonda, C. S. (2010). The wiley-blackwell handbook of infant development. In J. G. Bremner & T. D. Wachs (Eds.), (pp. 458–482). Wiley-Blackwell.

Botha, J. A., & Blunsom, P. (2013). Adaptor grammars for learning non-concatenative morphology. In *Emnlp* (pp. 345–356).

Boves, L., Ten Bosch, L., & Moore, R. K. (2007). ACORNS- Towards computational modeling of communication and recognition skills. In *6th IEEE International Conference on In Cognitive Informatics* (pp. 349–356). IEEE.

Brent, M. R. (1996a). Advances in the computational study of language acquisition. *Cognition*, *61*(1), 1–38.

Brent, M. R. (1996b). *Computational approaches to language acquisition*. MIT Press.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*(1-3), 71–105.

Brown, R. (1973). *A first language; the early stages*. Cambridge, Mass: Harvard University Press.

Bruner, J. S. (1975, April). The ontogenesis of speech acts. *Journal of Child Language*, *2*(01).

Bruner, J. S. (1983). *Child's Talk: Learning to Use Language*. New York, N.Y.: Norton.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., … DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *arXiv preprint arXiv:1406.3284*.

Carlin, M. A., Thomas, S., Jansen, A., & Hermansky, H. (2011). Rapid evaluation of speech representations for spoken term discovery. In *Proceedings of Interspeech.*

Caruana, R. (1997). Multitask learning. *Machine learning*, *28*(1), 41–75.

Casillas, M. (2016). Age and turn type in mayan children's predictions about conversational turn-taking. to be presented at. In *Boston university child language development.* Boston, USA.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT press.

Chouinard, M. M., & Clark, E. V. (2003). Adult reformulations of child errors as negative evidence. *Journal of child language*, *30*(03), 637–669.

Christiansen, M. H., Conway, C. M., & Curtin, S. (2005). Multiple-cue integration in language acquisition: A connectionist model of speech segmentation and rule-like behavior. *Language acquisition, change and emergence: Essay in evolutionary linguistics*, 205–249.

Christodoulopoulos, C., Goldwater, S., & Steedman, M. (2010). Two Decades of Unsupervised POS induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 575–584). Association for Computational Linguistics.

Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2008). Bootstrapping lexical and syntactic acquisition. *Language and Speech*, *51*(1-2), 61–75.

Clark, A., Giorgolo, G., & Lappin, S. (2013). Statistical representation of grammaticality judgements: the limits of n-gram models. In *Proceedings of the fourth annual workshop on cognitive modeling and computational linguistics (cmcl)* (pp. 28–36).

Clark, A., & Lappin, S. (2011). *Linguistic Nativism and the poverty of the stimulus*. Wiley and sons.

Connor, M., Fisher, C., & Roth, D. (2013). Starting from scratch in semantic role labeling: Early indirect supervision. In T. Poibeau, A. Villavicencio, A. Korhonen, & A. Alishahi (Eds.), *Cognitive aspects of computational language acquisition* (pp. 257–296). Springer.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Mit Press.

Daland, R., & Pierrehumbert, J. B. (2011). Learning Diphone-Based Segmentation. *Cognitive Science*, *35*(1), 119–155.

Dehaene-Lambertz, G., Dehaene, S., et al. (1994). Speed and cerebral correlates of syllable discrimination in infants. *Nature*, *370*(6487), 292–295.

de Marcken, C. G. (1996). *Unsupervised Language Acquisition* (Unpublished doctoral dissertation). MIT.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 248–255).

Devlin, J., Gupta, S., Girshick, R., Mitchell, M., & Zitnick, C. L. (2015). Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*.

Dresher, B. E., & Kaye, J. D. (1990). A computational learning model for metrical phonology. *Cognition*, *34*(2), 137–195.

D'Ulizia, A., Ferri, F., & Grifoni, P. (2011). A survey of grammatical inference methods for natural language learning. *Artificial Intelligence Review*, *36*(1), 1–27. doi: 10.1007/s10462-010-9199-1

Dunbar, E., Synnaeve, G., & Dupoux, E. (2015). On the origin of features: Quantitative methods for comparing representations. In *Abstract in glow-2015.*

Dupoux, E. (2016). *Evaluating models of language acquisition: are utility metrics useful?* Retrieved from `http://bootphon.blogspot.fr/2015/05/models-of-language-acquisition-machine.html`

Dwork, C. (2006). Differential privacy. In *Automata, languages and programming* (pp. 1–12). Springer.

Ehret, G., Munroe, P., Rice, K., & al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, *478*(7367), 103–9.

Eilers, R. E., Gavin, W., & Wilson, W. R. (1979). Linguistic experience and phonemic perception in infancy: A crosslinguistic study. *Child Development*, 14–18.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Elsner, M., Goldwater, S., & Eisenstein, J. (2012). Bootstrapping a Unified Model of Lexical and Phonetic Acquisition. In *Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics.*

Feldman, N., Myers, E., White, K., Griffiths, T., & Morgan, J. (2011). Learners use word-level statistics in phonetic category acquisition. In *Proceedings of the 35th Annual Boston University Conference on Language Development* (pp. 197–209).

Fernald, A. (2000). Speech to infants as hyperspeech: Knowledge-driven processes in early word recognition. *Phonetica*, *57*(2-4), 241–254.

Ferrucci, D. A. (2012). Introduction to "this is watson". *IBM Journal of Research and Development*, *56*(3.4), 1–1.

Fiscus, J. G., Ajot, J., Garofolo, J. S., & Doddingtion, G. (2007). Results of the 2006 spoken term detection evaluation. In *Proc. sigir* (Vol. 7, pp. 51–57).

Foppolo, F., Guasti, M. T., & Chierchia, G. (2012). Scalar implicatures in child language: Give children a chance. *Language learning and development*, *8*(4), 365–394.

Fourtassi, A., Boerschinger, B., Johnson, M., & Dupoux, E. (2013). Whyisenglishsoeasytosegment. In *Proceedings of the 4th workshop on cognitive modeling and computational linguistics (cmcl 2013)* (p. 1-10). Sofia, Bulgaria.

Fourtassi, A., & Dupoux, E. (2014). A rudimentary lexicon and semantics help bootstrap phoneme acquisition. In *Proceedings of the 18th conference on computational natural language learning (conll).*

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(3), 816.

Frank, M. (2015, December). *The manybabies project.* Retrieved from `http://babieslearninglanguage.blogspot.fr/2015/12/the-manybabies-project.html`

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010, November). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Fromkin, V. A. (1984). *Speech errors as linguistic evidence*. Walter de Gruyter.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*(3), 407–454.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2016). Region-based convolutional networks for accurate object detection and segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *38*(1), 142–158.

Gladkova, A., Drozd, A., Center, C., & Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of naacl-hlt* (pp. 8–15).

Glass, J. (2012). Towards unsupervised speech processing. In *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on* (pp. 1–4). IEEE.

Gold, E. M. (1967). Language identification in the limit. *Information and control*, *10*(5), 447–474.

Goldin-Meadow, S. (2005). *Hearing Gesture: How Our Hands Help Us Think.* Belknap Press of Harvard University Press.

Goldstein, M. H. (2008). Social Feedback to Babbling Facilitates Vocal Learning Michael H. Goldstein and Jennifer A. Schwade. *Psychological Science*.

Goldwater, S. J. (2007). *Nonparametric Bayesian models of lexical acquisition* (Unpublished doctoral dissertation). Brown.

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of child language*, *14*(01), 23–45.

Gregory, R. J. (2004). *Psychological testing: History, principles, and applications.* Allyn & Bacon.

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006, November). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, *30*(6), 509–516. doi: 10.1177/0165025406071914

Guenther, F. H., & Vladusich, T. (2012, September). A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, *25*(5), 408–422. doi: 10.1016/j.jneuroling.2009.08.006

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children.* Paul H Brookes Publishing.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *science*, *298*(5598), 1569–1579.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, *39*(3), 379–440.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the ieee international conference on computer vision* (pp. 1026–1034).

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099–3111.

Hirsh-Pasek, K., Nelson, D. G. K., Jusczyk, P. W., Cassidy, K. W., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition*, *26*(3), 269–286.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child development*, *74*(5), 1368–1378.

Hoff, E. (Ed.). (2012). *Research methods in child language: a practical guide*. Malden, MA: Wiley-Blackwell.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., … Bloom, L. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the society for research in child development*, i–135.

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive psychology*, *61*(4), 343–365.

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*(6), 645–654.

Isik, L., Tacchetti, A., & Poggio, T. (2016). Fast, invariant representation for human action in the visual system. *arXiv preprint arXiv:1601.01358*.

Jackendoff, R. (1997). *The architecture of the language faculty* (No. 28). MIT Press.

Jäger, G., & Rogers, J. (2012). Formal language theory: refining the chomsky hierarchy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1598), 1956–1970.

Jansen, A., Dupoux, E., Goldwater, S., Johnson, M., Khudanpur, S., Church, K., … others (2013). A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition. In *ICASSP* (pp. 8111–8115).

Jansen, A., Thomas, S., & Hermansky, H. (2013). Weak top-down constraints for unsupervised acoustic model training. In *ICASSP* (pp. 8091–8095).

Jelinek, F. (1997). *Statistical methods for speech recognition*. MIT press.

Johnson, K. (2004). Gold's theorem and cognitive science*. *Philosophy of Science*, *71*(4), 571–592.

Johnson, M. (2008). Using Adaptor Grammars to Identify Synergies in the Unsupervised Acquisition of Linguistic Structure. In *ACL* (pp. 398–406).

Johnson, W., & Reimers, P. (2010). *Patterns in child phonology*. Edinburgh University Press.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, Mass.: MIT Press.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology*, *29*(1), 1–23.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of memory and language*, *32*(3), 402–420.

Jusczyk, P. W., Hirsh-Pasek, K., Nelson, D. G. K., Kennedy, L. J., Woodward, A., & Piwoz, J. (1992). Perception of acoustic correlates of major phrasal units by young infants. *Cognitive psychology*, *24*(2), 252–293.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive psychology*, *39*(3), 159–207.

Kahou, S. E., Bouthillier, X., Lamblin, P., Gulcehre, C., Michalski, V., Konda, K., … others (2015). Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 1–13.

Katsos, N., & Bishop, D. V. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, *120*(1), 67–81.

Kelley, K. (1967). *Early syntactic acquisition* (Tech. Rep. No. P-3719). Santa Monica, California: Rand Corp.

Kohonen, T. (1988). The 'neural' phonetic typewriter. *Computer*, *21*(3), 11–22.

Kuhl, P. K. (1997, August). Cross-Language Analysis of Phonetic Units in Language Addressed to Infants. *Science*, *277*(5326), 684–686. doi: 10.1126/science.277.5326.684

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008, March). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000. doi: 10.1098/rstb.2007.2154

Kunkel, S., Schmidt, M., Eppler, J. M., Plesser, H. E., Masumoto, G., Igarashi, J., … others (2014). Spiking network simulation code for petascale computers. *Frontiers in neuroinformatics*, *8*(78).

Kwiatkowski, T., Goldwater, S., Zettlemoyer, L., & Steedman, M. (2012). A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. *EACL 2012*, 234.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Langley, P., & Carbonell, J. G. (1987). Language acquisition and machine learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 115–155). Hillsdale, N.J.: Lawrence Erlbaum.

Lee, C.-y., & Glass, J. (2012). A nonparametric Bayesian approach to acoustic model discovery. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 40–49).

Lee, C.-y., O'Donnell, T. J., & Glass, J. (2015). Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, *3*, 389–403.

Leibo, J. Z., Liao, Q., Anselmi, F., & Poggio, T. (2015). The invariance hypothesis implies domain-specific regions in visual cortex. *PLoS Comput Biol*, *11*(10), e1004390.

Levesque, H. J., Davis, E., & Morgenstern, L. (2011). The winograd schema challenge. In *Aaai spring symposium: Logical formalizations of commonsense reasoning.*

Lidz, J., & Musolino, J. (2002). Children's command of quantification. *Cognition*, *84*(2), 113–154.

Linzen, T., Dupoux, E., & Spector, B. (2016). Quantificational features in distributional word representations. In *Proceedings of the fifth joint conference on lexical and computational semantics (* sem 2016).*

Lu, C., & Tang, X. (2014). Surpassing human-level face verification performance on lfw with gaussianface. *arXiv preprint arXiv:1404.3840.*

Ludusan, B., Gravier, G., & Dupoux, E. (2014). Incorporating Prosodic Boundaries in Unsupervised Term Discovery. In *Proc. of Speech Prosody.*

Ludusan, B., Seidl, A., Dupoux, E., & Cristia, A. (2015). Motif discovery in infant-and adult-directed speech. In *Conference on empirical methods in natural language processing (emnlp)* (p. 93).

Ludusan, B., Versteegh, M., Jansen, A., Gravier, G., Cao, X.-N., Johnson, M., & Dupoux, E. (2014). Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems. In *Proceedings of LREC.*

MacWhinney, B. (1978). Conditions on acquisitional models. In *Proceedings of the ACM annual conference* (pp. 421–427). ACM.

MacWhinney, B. (1987). The Competition model. In B. MacWhinney (Ed.), (pp. 249–308). Hillsdale, N.J.: Lawrence Erlbaum.

MacWhinney, B. (2000). The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics*, *26*(4), 657–657.

Magri, G. (2015). Noise robustness and stochastic tolerance of OT error-driven ranking algorithms. *Journal of Logic and Computation.*

Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science*, *6*(5), 314.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, *46*(1), 53–85.

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, *37*, 103-124.

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants: A comprehensive test of the hyperarticulation hypothesis. *Psychological Science*, *26*(3), 341-347.

Martin, J. H., & Jurafsky, D. (2008). *Speech and language processing* (2nd ed. ed.). Pearson Prentice Hall.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

Mazuka, R., Cao, Y., Dupoux, E., & Christophe, A. (2011). The development of a phonological illusion: a cross-linguistic study with japanese and french infants. *Developmental science*, *14*(4), 693–699.

Mazuka, R., Igarashi, Y., & Nishikawa, K. (2006). *Input for learning japanese: Riken japanese mother-infant conversation corpus* (Vol. 106(165); Tech. Rep. No. TL 2006-16).

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, *12*(3), 369–378.

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchron, W. (2013, November). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition*, *129*(2), 362–378.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Morgan, J., & Demuth, K. (1996). *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition.* L. Erlbaum Associates.

Muscariello, A., Gravier, G., & Bimbot, F. (2009). Audio keyword extraction by unsupervised word discovery. In *INTERSPEECH 2009: 10th Annual Conference of the International Speech Communication Association.*

Mustafa, A., Kim, H., Guillemaut, J.-Y., & Hilton, A. (2016). Temporally coherent 4d reconstruction of complex dynamic scenes. *arXiv preprint arXiv:1603.03381.*

Ngon, C., Martin, A., Dupoux, E., Cabrol, D., Dutat, M., & Peperkamp, S. (2013). (Non)words, (non)words, (non)words: evidence for a protolexicon during the first year of life. *Developmental Science*, *16*(1), 24–34.

Olivier, D. C. (1968). *Stochastic grammars and language acquisition mechanisms* (Unpublished doctoral dissertation). Harvard University Doctoral dissertation.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Oudeyer, P.-Y., Kaplan, F., & Hafner, V. V. (2007, April). Intrinsic Motivation Systems for Autonomous Mental Development. *IEEE Transactions on Evolutionary Computation*, *11*(2), 265–286.

Pan, B. A., Rowe, M. L., Singer, J. D., & Snow, C. E. (2005). Maternal correlates of growth in toddler vocabulary production in low-income families. *Child development*, *76*(4), 763–782.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5206–5210).

Park, A. S., & Glass, J. R. (2008, January). Unsupervised Pattern Discovery in Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(1), 186–197.

Pearl, L., Goldwater, S., & Steyvers, M. (2010, September). Online Learning Mechanisms for Bayesian Models of Word Segmentation. *Research on Language and Computation*, *8*(2-3), 107–132. doi: 10.1007/s11168-011-9074-5

Pearl, L., & Phillips, L. (2016). Language, cognition, and computational models. In A. Villavicencio & T. Poibeau (Eds.), (chap. Evaluating language acquisition models: A utility-based look at

Bayesian segmentation). Cambridge Univ Press.

Peters, A. M. (1983). *The units of language acquisition* (Vol. 1). Cambridge University Press Archive.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, Mass: Harvard University Press.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 399–441). Lawrence Erlbaum.

Pinker, S. (1989). *Learnability and cognition: the acquisition of argument structure*. MIT Press.

Pinker, S. (1994). *The language instinct*. Harper.

Plummer, A., R. (2012). Aligning manifolds to model the earliest phonological abstraction in infant-caretaker vocal imitation. In *Interspeech* (pp. 2482–2485).

Poizner, H., Klima, E., & Bellugi, U. (1987). *What the hand reveals about the brain.* MIT Press Cambridge, MA.

Pons, C. G., Anguera, X., & Binefa, X. (2013). Two-Level Clustering towards Unsupervised Discovery of Acoustic Classes. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on* (Vol. 2, pp. 299–302). IEEE.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., . . . others (2011). The kaldi speech recognition toolkit. In *Ieee 2011 workshop on automatic speech recognition and understanding.*

Prins, R., & Van Noord, G. (2001). Unsupervised pos-tagging improves parsing accuracy and parsing efficiency. In *Iwpt.*

Rahmani, H., Mian, A., & Shah, M. (2016). Learning a deep model for human action recognition from novel viewpoints. *arXiv preprint arXiv:1602.00828.*

Räsänen, O. (2012). Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Communication*, *54*(9), 975–997. doi: 10.1016/j.specom.2012.05.001

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914.

Rowe, M. L., & Goldin-Meadow, S. (2009). Differences in early gesture explain ses disparities in child vocabulary size at school entry. *Science*, *323*(5916), 951–953.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668.

Roy, D. (2009). New horizons in the study of child language acquisition. In *Proceedings of interspeech.* Brighton, England.

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive science*, *26*(1), 113–146.

Rumelhart, D. E., & McClelland, J. L. (1987). Mechanisms of language acquisition. In B. MacWhinney (Ed.), (pp. 195–248). Erlbaum Hillsdale, NJ.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Sakakibara, Y. (1992). Efficient learning of context-free grammars from positive structural examples. *Information and Computation*, *97*(1), 23–60.

Sakas, W. G., & Fodor, J. D. (2012). Disambiguating syntactic triggers. *Language Acquisition*, *19*(2), 83–143.

Sangwan, A., Hansen, J., Irvin, D., Crutchfield, S., & Greenwood, C. (2015). Studying the relationship between physical and language environments of children: Who's speaking to whom and where? In *Signal processing and signal processing education workshop (sp/spe), 2015 ieee* (pp. 49–54).

Saxton, M. (1997). The contrast theory of negative input. *Journal of child language*, *24*(01), 139–161.

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline. In *INTERSPEECH-2013* (p. 1781-1785). Lyon, France.

Shneidman, L. A., Arroyo, M. E., Levine, S. C., & Goldin-Meadow, S. (2013, June). What counts as effective input for word learning? *Journal of Child Language*, *40*(03), 672–686.

Shneidman, L. A., & Goldin-Meadow, S. (2012, September). Language input and acquisition in a Mayan village: how important is directed speech?: Mayan village. *Developmental Science*, *15*(5), 659–673.

Siklossy, L. (1968). *Natural language learning by computer* (Tech. Rep.). DTIC Document.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., . . . others (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1), 39–91.

Siu, M.-h., Gish, H., Chan, A., Belfield, W., & Lowe, S. (2013). Unsupervized training of an HMM-based self-organizing recognizer with applications to topic classification and keyword discovery. *Computer Speech & Language*.

Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480–498.

Smith, L. B., Yu, C., Yoshida, H., & Fausey, C. M. (2015). Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, *16*(3), 407–419.

Smith, N. A., & Trainor, L. J. (2008). Infant-directed speech is modulated by infant feedback. *Infancy*, *13*(4), 410–420.

Snow, C. E. (1972, June). Mothers' Speech to Children Learning Language. *Child Development*, *43*(2), 549.

Song, J. J. (2010). *The oxford handbook of linguistic typology*. Oxford Univ. Press.

Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, *134*, 219–248.

Stark, R. (1980). Child phonology. Vol. 1: Production. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), (chap. Stages of development in the first year of life). New York: Acad. Press.

Steedman, M. (2014). Evolutionary basis for human language: Comment on "Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition" by tecumseh fitch. *Physics of life reviews*, *11*(3), 382–388.

Sundara, M., Polka, L., & Genesee, F. (2006). Language-experience facilitates discrimination of/d-/in monolingual and bilingual acquisition of english. *Cognition*, *100*(2), 369–388.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Tamis-LeMonda, C. S., & Rodriguez, E. T. (2008). Parents' role in fostering young children's learning and language development. In (pp. 1–11).

Ten Bosch, L., & Cranen, B. (2007). A computational model for unsupervised word discovery. In *INTERSPEECH* (pp. 1481–1484).

Tesar, B., & Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry*, *29*(2), 229–268.

Tesar, B., & Smolensky, P. (2000). *Learnability in optimality theory*. Mit Press.

Thiollière, R., Dunbar, E., Synnaeve, G., Versteegh, M., & Dupoux, E. (2015). A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. In *INTERSPEECH-2015*.

Thomas, D. G., Campos, J. J., Shucard, D. W., Ramsay, D. S., & Shucard, J. (1981). Semantic comprehension in infancy: A signal detection analysis. *Child development*, 798–803.

Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Cambridge, Mass: Harvard University Press.

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses toward cumulative data assessment. *Perspectives on Psychological Science*, *9*(6), 661–665.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273–13278.

Van Cleve, J. V. (2004). *Genetics, disability, and deafness*. Gallaudet University Press.

VanDam, M., Warlaumont, A. S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., & MacWhinney, B. (2016). Homebank: An online repository of daylong child-centered audio recordings. In *Seminars in speech and language* (Vol. 37, pp. 128–142).

van de Weijer, J. (2002). How much does an infant hear in a day. In *GALA 2001 Conference on Language Acquisition, Lisboa.*

Varadarajan, B., Khudanpur, S., & Dupoux, E. (2008). Unsupervised learning of acoustic subword units. In *Proceedings of acl-08: Hlt* (p. 165-168).

Versteegh, M., Anguera, X., Jansen, A., & Dupoux, E. (2016). The zero resource speech challenge 2015: Proposed approaches and results. In *SLTU-2016.*

Versteegh, M., Thiollière, R., Schatz, T., Cao, X.-N., Anguera, X., Jansen, A., & Dupoux, E. (2015). The zero resource speech challenge 2015. In *INTERSPEECH-2015.*

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. *Psychological science*, *25*(7), 1314-1324.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392–393.

Weisleder, A., & Fernald, A. (2013, November). Talking to Children Matters: Early Language Experience Strengthens Processing and Builds Vocabulary. *Psychological Science*, *24*(11), 2143–2152.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language learning and development*, *1*(2), 197–234.

Werker, J. F., & Tees, R. C. (1984). Cross-language Speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.

Word2vec google project page. (2013, july). Retrieved from `https://code.google.com/archive/p/word2vec/`

Xu, D., Yapanel, U. H., Gray, S. S., Gilkerson, J., Richards, J. A., & Hansen, J. H. (2008). Signal processing for young child speech language development. In *Wocci* (p. 20).

Yang, C. D. (2002). *Knowledge and learning in natural language*. Oxford University Press.

Yu, C., & Smith, A. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies

Tal Linzen[1,2] Emmanuel Dupoux[1]
LSCP[1] & IJN[2], CNRS,
EHESS and ENS, PSL Research University
{tal.linzen,
emmanuel.dupoux}@ens.fr

Yoav Goldberg
Computer Science Department
Bar Ilan University
yoav.goldberg@gmail.com

Abstract

The success of long short-term memory (LSTM) neural networks in language processing is typically attributed to their ability to capture long-distance statistical regularities. Linguistic regularities are often sensitive to syntactic structure; can such dependencies be captured by LSTMs, which do not have explicit structural representations? We begin addressing this question using number agreement in English subject-verb dependencies. We probe the architecture's grammatical competence both using training objectives with an explicit grammatical target (number prediction, grammaticality judgments) and using language models. In the strongly supervised settings, the LSTM achieved very high overall accuracy (less than 1% errors), but errors increased when sequential and structural information conflicted. The frequency of such errors rose sharply in the language-modeling setting. We conclude that LSTMs can capture a non-trivial amount of grammatical structure given targeted supervision, but stronger architectures may be required to further reduce errors; furthermore, the language modeling signal is insufficient for capturing syntax-sensitive dependencies, and should be supplemented with more direct supervision if such dependencies need to be captured.

1 Introduction

Recurrent neural networks (RNNs) are highly effective models of sequential data (Elman, 1990). The rapid adoption of RNNs in NLP systems in recent years, in particular of RNNs with gating mechanisms such as long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997) or gated recurrent units (GRU) (Cho et al., 2014), has led to significant gains in language modeling (Mikolov et al., 2010; Sundermeyer et al., 2012), parsing (Vinyals et al., 2015; Kiperwasser and Goldberg, 2016; Dyer et al., 2016), machine translation (Bahdanau et al., 2015) and other tasks.

The effectiveness of RNNs[1] is attributed to their ability to capture statistical contingencies that may span an arbitrary number of words. The word *France*, for example, is more likely to occur somewhere in a sentence that begins with *Paris* than in a sentence that begins with *Penguins*. The fact that an arbitrary number of words can intervene between the mutually predictive words implies that they cannot be captured by models with a fixed window such as n-gram models, but can in principle be captured by RNNs, which do not have an architecturally fixed limit on dependency length.

RNNs are sequence models: they do not explicitly incorporate syntactic structure. Indeed, many word co-occurrence statistics can be captured by treating the sentence as an unstructured list of words (*Paris-France*); it is therefore unsurprising that RNNs can learn them well. Other dependencies, however, are sensitive to the syntactic structure of the sentence (Chomsky, 1965; Everaert et al., 2015). To what extent can RNNs learn to model such phenomena based only on sequential cues?

Previous research has shown that RNNs (in particular LSTMs) can learn artificial context-free languages (Gers and Schmidhuber, 2001) as well as nesting and

[1] In this work we use the term RNN to refer to the entire class of sequential recurrent neural networks. Instances of the class include long short-term memory networks (LSTM) and the Simple Recurrent Network (SRN) due to Elman (1990).

Transactions of the Association for Computational Linguistics, vol. 4, pp. 521–535, 2016. Action Editor: Hinrich Schütze.
Submission batch: 1/2016; Revision batch: 9/2016; Published 12/2016.

indentation in a programming language (Karpathy et al., 2016). The goal of the present work is to probe their ability to learn *natural language* hierarchical (syntactic) structures from a corpus without syntactic annotations. As a first step, we focus on a particular dependency that is commonly regarded as evidence for hierarchical structure in human language: English subject-verb agreement, the phenomenon in which the form of a verb depends on whether the subject is singular or plural (*the kids play* but *the kid plays*; see additional details in Section 2). If an RNN-based model succeeded in learning this dependency, that would indicate that it can learn to approximate or even faithfully implement syntactic structure.

Our main interest is in whether LSTMs have the *capacity* to learn structural dependencies from a natural corpus. We therefore begin by addressing this question under the most favorable conditions: training with explicit supervision. In the setting with the strongest supervision, which we refer to as the number prediction task, we train it directly on the task of guessing the number of a verb based on the words that preceded it (Sections 3 and 4). We further experiment with a grammaticality judgment training objective, in which we provide the model with full sentences annotated as to whether or not they violate subject-verb number agreement, without an indication of the locus of the violation (Section 5). Finally, we trained the model without any grammatical supervision, using a language modeling objective (predicting the next word).

Our quantitative results (Section 4) and qualitative analysis (Section 7) indicate that most naturally occurring agreement cases in the Wikipedia corpus are easy: they can be resolved without syntactic information, based only on the sequence of nouns preceding the verb. This leads to high overall accuracy in all models. Most of our experiments focus on the supervised number prediction model. The accuracy of this model was lower on harder cases, which require the model to encode or approximate structural information; nevertheless, it succeeded in recovering the majority of agreement cases even when four nouns of the opposite number intervened between the subject and the verb (17% errors). Baseline models failed spectacularly on these hard cases, performing far below chance levels. Fine-grained analysis revealed that mistakes are much more common when no overt cues to syntactic structure (in particular function words) are available, as is the case in noun-noun compounds and reduced relative clauses. This indicates that the number prediction model indeed managed to capture a decent amount of syntactic knowledge, but was overly reliant on function words.

Error rates increased only mildly when we switched to more indirect supervision consisting only of sentence-level grammaticality annotations without an indication of the crucial verb. By contrast, the language model trained without explicit grammatical supervision performed worse than chance on the harder agreement prediction cases. Even a state-of-the-art large-scale language model (Jozefowicz et al., 2016) was highly sensitive to recent but structurally irrelevant nouns, making more than five times as many mistakes as the number prediction model on these harder cases. These results suggest that explicit supervision is necessary for learning the agreement dependency using this architecture, limiting its plausibility as a model of child language acquisition (Elman, 1990). From a more applied perspective, this result suggests that for tasks in which it is desirable to capture syntactic dependencies (e.g., machine translation or language generation), language modeling objectives should be supplemented by supervision signals that directly capture the desired behavior.

2 Background: Subject-Verb Agreement as Evidence for Syntactic Structure

The form of an English third-person present tense verb depends on whether the head of the *syntactic subject* is plural or singular:[2]

(1) a. The **key is** on the table.
b. *The **key are** on the table.
c. *The **keys is** on the table.
d. The **keys are** on the table.

While in these examples the subject's head is adjacent to the verb, in general the two can be separated by some sentential material:[3]

[2] Identifying the head of the subject is typically straightforward. In what follows we will use the shorthand "the subject" to refer to the head of the subject.

[3] In the examples, the subject and the corresponding verb are marked in boldface, agreement attractors are underlined and intervening nouns of the same number as the subject are marked in italics. Asterisks mark unacceptable sentences.

(2) The **keys** to the cabinet **are** on the table.

Given a syntactic parse of the sentence and a verb, it is straightforward to identify the head of the subject that corresponds to that verb, and use that information to determine the number of the verb (Figure 1).

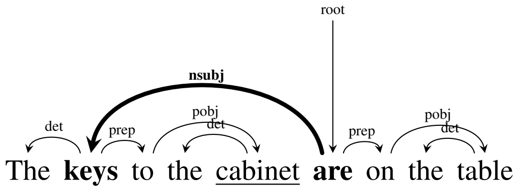

Figure 1: The form of the verb is determined by the head of the subject, which is directly connected to it via an *nsubj* edge. Other nouns that intervene between the head of the subject and the verb (here *cabinet* is such a noun) are irrelevant for determining the form of the verb and need to be ignored.

By contrast, models that are insensitive to structure may run into substantial difficulties capturing this dependency. One potential issue is that there is no limit to the complexity of the subject NP, and any number of sentence-level modifiers and parentheticals—and therefore an arbitrary number of words—can appear between the subject and the verb:

(3) The **building** on the far right that's quite old and run down **is** the Kilgore Bank Building.

This property of the dependency entails that it cannot be captured by an n-gram model with a fixed n. RNNs are in principle able to capture dependencies of an unbounded length; however, it is an empirical question whether or not they will learn to do so in practice when trained on a natural corpus.

A more fundamental challenge that the dependency poses for structure-insensitive models is the possibility of *agreement attraction errors* (Bock and Miller, 1991). The correct form in (3) could be selected using simple heuristics such as "agree with the most recent noun", which are readily available to sequence models. In general, however, such heuristics are unreliable, since other nouns can intervene between the subject and the verb in the linear sequence of the sentence. Those intervening nouns can have the same number as the subject, as in (4), or the opposite number as in (5)-(7):

(4) Alluvial **soils** carried in the *floodwaters* **add** nutrients to the floodplains.

(5) The only championship **banners** that are currently displayed within the building **are** for national or NCAA Championships.

(6) The **length** of the forewings **is** 12-13.

(7) Yet the **ratio** of men who survive to the women and children who survive **is** not clear in this story.

Intervening nouns with the opposite number from the subject are called *agreement attractors*. The potential presence of agreement attractors entails that the model must identify the head of the syntactic subject that corresponds to a given verb in order to choose the correct inflected form of that verb.

Given the difficulty in identifying the subject from the linear sequence of the sentence, dependencies such as subject-verb agreement serve as an argument for structured syntactic representations in humans (Everaert et al., 2015); they may challenge models such as RNNs that do not have pre-wired syntactic representations. We note that subject-verb number agreement is only one of a number of structure-sensitive dependencies; other examples include negative polarity items (e.g., *any*) and reflexive pronouns (*herself*). Nonetheless, a model's success in learning subject-verb agreement would be highly suggestive of its ability to master hierarchical structure.

3 The Number Prediction Task

To what extent can a sequence model learn to be sensitive to the hierarchical structure of natural language? To study this question, we propose the *number prediction* task. In this task, the model sees the sentence up to but not including a present-tense verb, e.g.:

(8) The keys to the cabinet ____

It then needs to guess the number of the following verb (a binary choice, either PLURAL or SINGULAR). We examine variations on this task in Section 5.

In order to perform well on this task, the model needs to encode the concepts of *syntactic number* and *syntactic subjecthood*: it needs to learn that some words are singular and others are plural, and to be able to identify the correct subject. As we have illustrated in Section 2, correctly identifying the subject that corresponds to a particular verb often requires sensitivity to hierarchical syntax.

Data: An appealing property of the number prediction task is that we can generate practically unlimited training and testing examples for this task by querying a corpus for sentences with present-tense verbs, and noting the number of the verb. Importantly, we do not need to correctly identify the subject in order to create a training or test example. We generated a corpus of ∼1.35 million number prediction problems based on Wikipedia, of which ∼121,500 (9%) were used for training, ∼13,500 (1%) for validation, and the remaining ∼1.21 million (90%) were reserved for testing.[4] The large number of test sentences was necessary to ensure that we had a good variety of test sentences representing less common constructions (see Section 4).[5]

Model and baselines: We encode words as one-hot vectors: the model does not have access to the characters that make up the word. Those vectors are then embedded into a 50-dimensional vector space. An LSTM with 50 hidden units reads those embedding vectors in sequence; the state of the LSTM at the end of the sequence is then fed into a logistic regression classifier. The network is trained[6] in an end-to-end fashion, including the word embeddings.[7]

To isolate the effect of syntactic structure, we also consider a baseline which is exposed only to the nouns in the sentence, in the order in which they appeared originally, and is then asked to predict the number of the following verb. The goal of this baseline is to withhold the syntactic information carried by function words, verbs and other parts of speech. We explore two variations on this baseline: one that only receives common nouns (*dogs*, *pipe*), and another that also receives pronouns (*he*) and proper nouns (*France*). We refer to these as the *noun-only baselines*.

4 Number Prediction Results

Overall accuracy: Accuracy was very high overall: the system made an incorrect number prediction only in 0.83% of the dependencies. The noun-only baselines performed significantly worse: 4.2% errors for the common-nouns case and 4.5% errors for the all-nouns case. This suggests that function words, verbs and other syntactically informative elements play an important role in the model's ability to correctly predict the verb's number. However, while the noun-only baselines made more than four times as many mistakes as the number prediction system, their still-low absolute error rate indicates that around 95% of agreement dependencies can be captured based solely on the sequence of nouns preceding the verb. This is perhaps unsurprising: sentences are often short and the verb is often directly adjacent to the subject, making the identification of the subject simple. To gain deeper insight into the syntactic capabilities of the model, then, the rest of this section investigates its performance on more challenging dependencies.[8]

Distance: We first examine whether the network shows evidence of generalizing to dependencies where the subject and the verb are far apart. We focus in this analysis on simpler cases where no nouns intervened between the subject and the verb. As Figure 2a shows, performance did not degrade considerably when the distance between the subject and the verb grew up to 15 words (there were very few longer dependencies). This indicates that the network generalized the dependency from the common distances of 0 and 1 to rare distances of 10 and more.

Agreement attractors: We next examine how the model's error rate was affected by nouns that intervened between the subject and the verb in the linear

[4] We limited our search to sentences that were shorter than 50 words. Whenever a sentence had more than one subject-verb dependency, we selected one of the dependencies at random.

[5] Code and data are available at `http://tallinzen.net/projects/lstm_agreement`.

[6] The network was optimized using Adam (Kingma and Ba, 2015) and early stopping based on validation set error. We trained the number prediction model 20 times with different random initializations, and report accuracy averaged across all runs. The models described in Sections 5 and 6 are based on 10 runs, with the exception of the language model, which is slower to train and was trained once.

[7] The size of the vocabulary was capped at 10000 (after lowercasing). Infrequent words were replaced with their part of speech (Penn Treebank tagset, which explicitly encodes number distinctions); this was the case for 9.6% of all tokens and 7.1% of the subjects.

[8] These properties of the dependencies were identified by parsing the test sentences using the parser described in Goldberg and Nivre (2012).

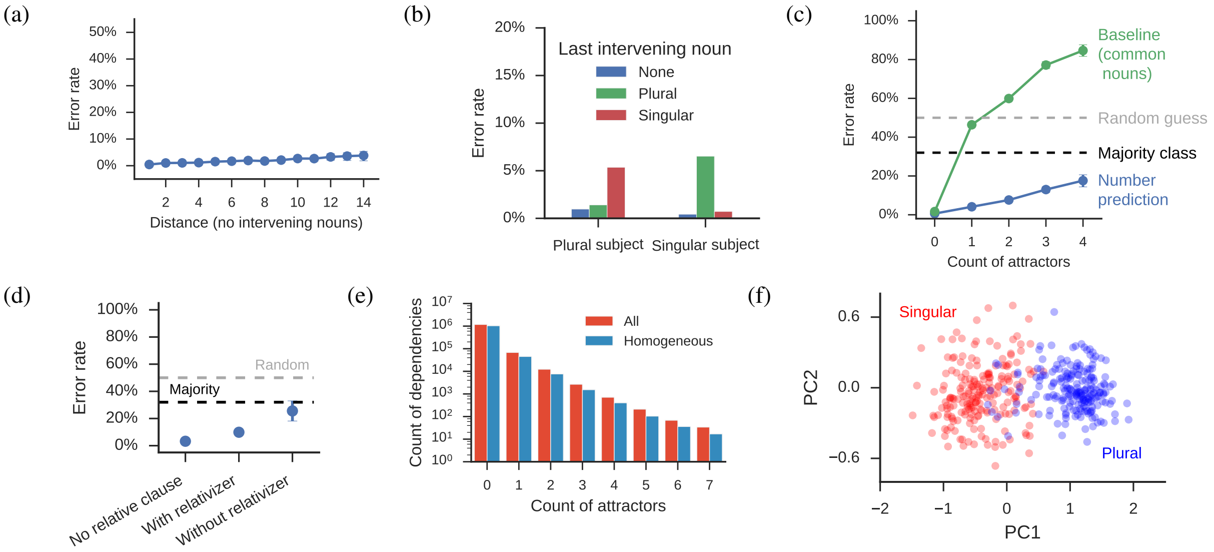

Figure 2: **(a-d)** Error rates of the LSTM number prediction model as a function of: (a) distance between the subject and the verb, in dependencies that have no intervening nouns; (b) presence and number of last intervening noun; (c) count of attractors in dependencies with homogeneous intervention; (d) presence of a relative clause with and without an overt relativizer in dependencies with homogeneous intervention and exactly one attractor. All error bars represent 95% binomial confidence intervals.

(e-f) Additional plots: (e) count of attractors per dependency in the corpus (note that the y-axis is on a log scale); (f) embeddings of singular and plural nouns, projected onto their first two principal components.

order of the sentence. We first focus on whether or not there were any intervening nouns, and if there were, whether the number of the subject differed from the number of the last intervening noun—the type of noun that would trip up the simple heuristic of agreeing with the most recent noun.

As Figure 2b shows, a last intervening noun of the same number as the subject increased error rates only moderately, from 0.4% to 0.7% in singular subjects and from 1% to 1.4% in plural subjects. On the other hand, when the last intervening noun was an agreement attractor, error rates increased by almost an order of magnitude (to 6.5% and 5.4% respectively). Note, however, that even an error rate of 6.5% is quite impressive considering uninformed strategies such as random guessing (50% error rate), always assigning the more common class label (32% error rate, since 32% of the subjects in our corpus are plural) and the number-of-most-recent-noun heuristic (100% error rate). The noun-only LSTM baselines performed much worse in agreement attraction cases, with error rates of 46.4% (common nouns) and 40% (all nouns).

We next tested whether the effect of attractors is cumulative, by focusing on dependencies with multiple attractors. To avoid cases in which the effect of an attractor is offset by an intervening noun with the same number as the subject, we restricted our search to dependencies in which all of the intervening nouns had the same number, which we term *dependencies with homogeneous intervention*. For example, (9) has homogeneous intervention whereas (10) does not:

(9) The **roses** in the vase by the door **are** red.

(10) The **roses** in the vase by the *chairs* **are** red.

Figure 2c shows that error rates increased gradually as more attractors intervened between the subject and the verb. Performance degraded quite slowly, however: even with four attractors the error rate was only 17.6%. As expected, the noun-only baselines performed significantly worse in this setting, reaching an error rate of up to 84% (worse than chance) in the case of four attractors. This confirms that syntactic cues are critical for solving the harder cases.

Relative clauses: We now look in greater detail into the network's performance when the words that intervened between the subject and verb contained a relative clause. Relative clauses with attractors are likely to be fairly challenging, for several reasons. They typically contain a verb that agrees with the attractor, reinforcing the misleading cue to noun number. The attractor is often itself a subject of an irrelevant verb, making a potential "agree with the most recent subject" strategy unreliable. Finally, the existence of a relative clause is sometimes not overtly indicated by a function word (relativizer), as in (11) (for comparison, see the minimally different (12)):

(11) The **landmarks** this <u>article</u> lists here **are** also run-of-the-mill and not notable.

(12) The **landmarks** *that* this <u>article</u> lists here **are** also run-of-the-mill and not notable.

For data sparsity reasons we restricted our attention to dependencies with a single attractor and no other intervening nouns. As Figure 2d shows, attraction errors were more frequent in dependencies with an overt relative clause (9.9% errors) than in dependencies without a relative clause (3.2%), and considerably more frequent when the relative clause was not introduced by an overt relativizer (25%). As in the case of multiple attractors, however, while the model struggled with the more difficult dependencies, its performance was much better than random guessing, and slightly better than a majority-class strategy.

Word representations: We explored the 50-dimensional word representations acquired by the model by performing a principal component analysis. We assigned a part-of-speech (POS) to each word based on the word's most common POS in the corpus. We only considered relatively unambiguous words, in which a single POS accounted for more than 90% of the word's occurrences in the corpus. Figure 2f shows that the first principal component corresponded almost perfectly to the expected number of the noun, suggesting that the model learned the number of specific words very well; recall that the model did not have access during training to noun number annotations or to morphological suffixes such as *-s* that could be used to identify plurals.

Visualizing the network's activations: We start investigating the inner workings of the number prediction network by analyzing its activation in response to particular syntactic constructions. To simplify the analysis, we deviate from our practice in the rest of this paper and use constructed sentences.

We first constructed sets of sentence prefixes based on the following patterns:

(13) **PP:** The toy(s) of the boy(s)...

(14) **RC:** The toy(s) that the boy(s)...

These patterns differ by exactly one function word, which determines the type of the modifier of the main clause subject: a prepositional phrase (PP) in the first sentence and a relative clause (RC) in the second. In PP sentences the correct number of the upcoming verb is determined by the main clause subject *toy(s)*; in RC sentences it is determined by the embedded subject *boy(s).*

We generated all four versions of each pattern, and repeated the process ten times with different lexical items (*the house(s) of/that the girl(s)*, *the computer(s) of/that the student(s)*, etc.), for a total of 80 sentences. The network made correct number predictions for all 40 PP sentences, but made three errors in RC sentences. We averaged the word-by-word activations across all sets of ten sentences that had the same combination of modifier (PP or RC), first noun number and second noun number. Plots of the activation of all 50 units are provided in the Appendix (Figure 5). Figure 3a highlights a unit (Unit 1) that shows a particularly clear pattern: it tracks the number of the main clause subject throughout the PP modifier; by contrast, it resets when it reaches the relativizer *that* which introduces the RC modifier, and then switches to tracking the number of the embedded subject.

To explore how the network deals with dependencies spanning a larger number of words, we tracked its activation during the processing of the following two sentences:[9]

(15) The houses of/that the man from the office across the street...

The network made the correct prediction for the PP

[9] We simplified this experiment in light of the relative robustness of the first experiment to lexical items and to whether each of the nouns was singular or plural.

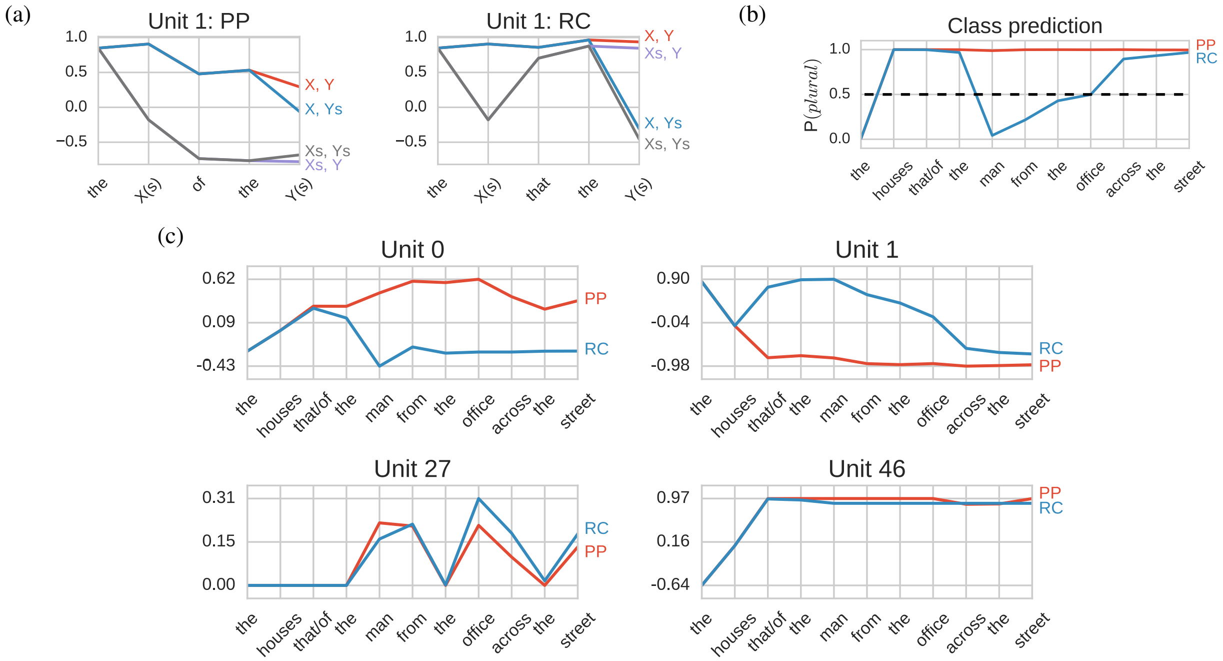

Figure 3: Word-by-word visualization of LSTM activation: (a) a unit that correctly predicts the number of an upcoming verb. This number is determined by the first noun (X) when the modifier is a prepositional phrase (PP) and by the second noun (Y) when it is an object relative clause (RC); (b) the evolution of the predictions in the case of a longer modifier: the predictions correctly diverge at the embedded noun, but then incorrectly converge again; (c) the activation of four representative units over the course of the same sentences.

but not the RC sentence (as before, the correct predictions are PLURAL for PP and SINGULAR for RC). Figure 3b shows that the network begins by making the correct prediction for RC immediately after *that*, but then falters: as the sentence goes on, the resetting effect of *that* diminishes. The activation time courses shown in Figure 3c illustrate that Unit 1, which identified the subject correctly when the prefix was short, gradually forgets that it is in an embedded clause as the prefix grows longer. By contrast, Unit 0 shows a stable capacity to remember the current embedding status. Additional representative units shown in Figure 3c are Unit 46, which consistently stores the number of the main clause subject, and Unit 27, which tracks the number of the most recent noun, resetting at noun phrase boundaries.

While the interpretability of these patterns is encouraging, our analysis only scratches the surface of the rich possibilities of a linguistically-informed analysis of a neural network trained to perform a syntax-sensitive task; we leave a more extensive investigation for future work.

5 Alternative Training Objectives

The number prediction task followed a fully supervised objective, in which the network identifies the number of an upcoming verb based only on the words preceding the verb. This section proposes three objectives that modify some of the goals and assumptions of the number prediction objective (see Table 1 for an overview).

Verb inflection: This objective is similar to number prediction, with one difference: the network receives not only the words leading up to the verb, but also the singular form of the upcoming verb (e.g., *writes*). In practice, then, the network needs to decide between the singular and plural forms of a particular verb (*writes* or *write*). Having access to the semantics of the verb can help the network identify the noun that serves as its subject without using the syntactic subjecthood criteria. For example, in the following sentence:

(16) People from the capital often eat pizza.

Training objective	Sample input	Training signal	Prediction task	Correct answer
Number prediction	*The keys to the cabinet*	PLURAL	SINGULAR/PLURAL?	PLURAL
Verb inflection	*The keys to the cabinet [is/are]*	PLURAL	SINGULAR/PLURAL?	PLURAL
Grammaticality	*The keys to the cabinet are here.*	GRAMMATICAL	GRAMMATICAL/UNGRAMMATICAL?	GRAMMATICAL
Language model	*The keys to the cabinet*	are	$P(are) > P(is)$?	True

Table 1: Examples of the four training objectives and corresponding prediction tasks.

only *people* is a plausible subject for *eat*; the network can use this information to infer that the correct form of the verb is *eat* is rather than *eats*.

This objective is similar to the task that humans face during language production: after the speaker has decided to use a particular verb (e.g., *write*), he or she needs to decide whether its form will be *write* or *writes* (Levelt et al., 1999; Staub, 2009).

Grammaticality judgments: The previous objectives explicitly indicate the location in the sentence in which a verb can appear, giving the network a cue to syntactic clause boundaries. They also explicitly direct the network's attention to the number of the verb. As a form of weaker supervision, we experimented with a grammaticality judgment objective. In this scenario, the network is given a complete sentence, and is asked to judge whether or not it is grammatical.

To train the network, we made half of the examples in our training corpus ungrammatical by flipping the number of the verb.[10] The network read the entire sentence and received a supervision signal at the end. This task is modeled after a common human data collection technique in linguistics (Schütze, 1996), although our training regime is of course very different to the training that humans are exposed to: humans rarely receive ungrammatical sentences labeled as such (Bowerman, 1988).

Language modeling (LM): Finally, we experimented with a word prediction objective, in which the model did not receive any grammatically relevant supervision (Elman, 1990; Elman, 1991). In this scenario, the goal of the network is to predict the next word at each point in every sentence. It receives unlabeled sentences and is not specifically instructed to attend to the number of the verb. In the network that implements this training scenario, RNN activation after each word is fed into a fully connected dense layer followed by a softmax layer over the entire vocabulary.

We evaluate the knowledge that the network has acquired about subject-verb noun agreement using a task similar to the verb inflection task. To perform the task, we compare the probabilities that the model assigns to the two forms of the verb that in fact occurred in the corpus (e.g., *write* and *writes*), and select the form with the higher probability.[11] As this task is not part of the network's training objective, and the model needs to allocate considerable resources to predicting each word in the sentence, we expect the LM to perform worse than the explicitly supervised objectives.

Results: When considering all agreement dependencies, all models achieved error rates below 7% (Figure 4a); as mentioned above, even the noun-only number prediction baselines achieved error rates below 5% on this task. At the same time, there were large differences in accuracy across training objectives. The verb inflection network performed slightly but significantly better than the number prediction one (0.8% compared to 0.83% errors), suggesting that the semantic information carried by the verb is moderately helpful. The grammaticality judgment objective performed somewhat worse, at 2.5% errors, but still outperformed the noun-only baselines by a large margin, showing the capacity of the LSTM architecture to learn syntactic dependencies even given fairly indirect evidence.

[10] In some sentences this will not in fact result in an ungrammatical sentence, e.g. with collective nouns such as *group*, which are compatible with both singular and plural verbs in some dialects of English (Huddleston and Pullum, 2002); those cases appear to be rare.

[11] One could also imagine performing the equivalent of the number prediction task by aggregating LM probability mass over all plural verbs and all singular verbs. This approach may be more severely affected by part-of-speech ambiguous words than the one we adopted; we leave the exploration of this approach to future work.

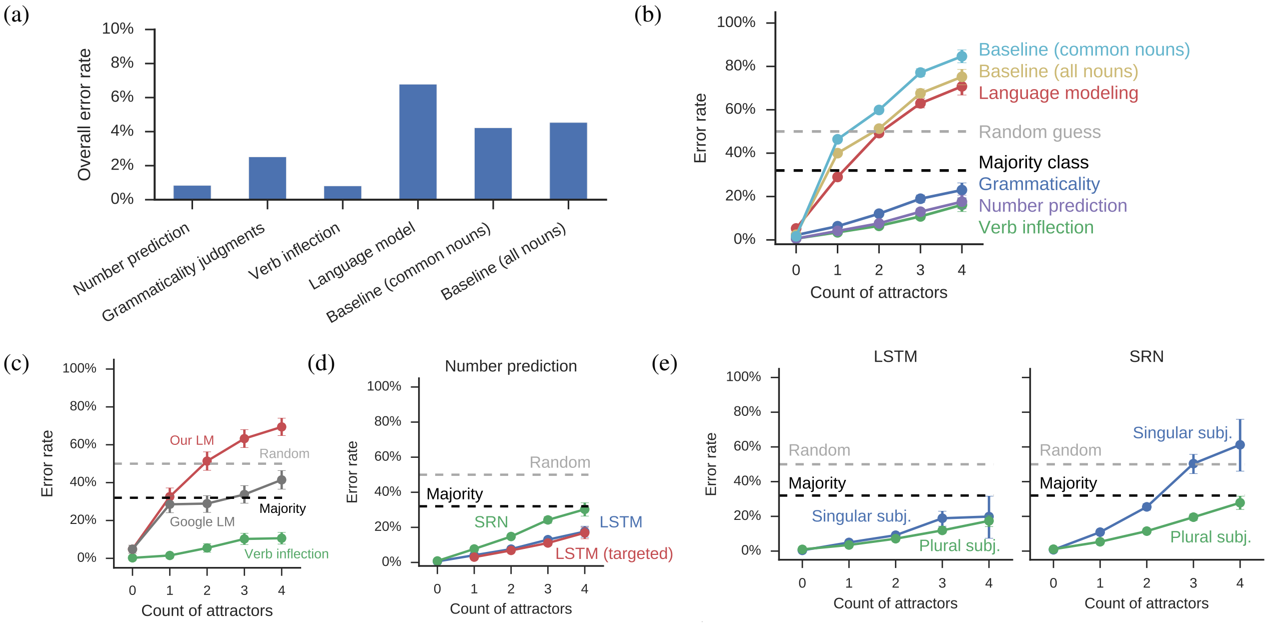

Figure 4: Alternative tasks and additional experiments: (a) overall error rate across tasks (note that the y-axis ends in 10%); (b) effect of count of attractors in homogeneous dependencies across training objectives; (c) comparison of the Google LM (Jozefowicz et al., 2016) to our LM and one of our supervised verb inflection systems, on a sample of sentences; (d) number prediction: effect of count of attractors using SRNs with standard training or LSTM with targeted training; (e) number prediction: difference in error rate between singular and plural subjects across RNN cell types. Error bars represent binomial 95% confidence intervals.

The worst performer was the language model. It made eight times as many errors as the original number prediction network (6.78% compared to 0.83%), and did substantially worse than the noun-only baselines (though recall that the noun-only baselines were still explicitly trained to predict verb number).

The differences across the networks are more striking when we focus on dependencies with agreement attractors (Figure 4b). Here, the language model does worse than chance in the most difficult cases, and only slightly better than the noun-only baselines. The worse-than-chance performance suggests that attractors actively confuse the networks rather than cause them to make a random decision. The other models degrade more gracefully with the number of agreement attractors; overall, the grammaticality judgment objective is somewhat more difficult than the number prediction and verb inflection ones. In summary, we conclude that while the LSTM is capable of learning syntax-sensitive agreement dependencies under various objectives, the language-modeling objective alone is not sufficient for learning such dependencies, and a more direct form of training signal is required.

Comparison to a large-scale language model: One objection to our language modeling result is that our LM faced a much harder objective than our other models—predicting a distribution over 10,000 vocabulary items is certainly harder than binary classification—but was equipped with the same capacity (50-dimensional hidden state and word vectors). Would the performance gap between the LM and the explicitly supervised models close if we increased the capacity of the LM?

We address this question using a very large publicly available LM (Jozefowicz et al., 2016), which we refer to as the Google LM.[12] The Google LM represents the current state-of-the-art in language modeling: it is trained on a billion-word corpus (Chelba et al., 2013), with a vocabulary of 800,000 words. It is based on a two-layer LSTM with 8192 units in each layer, or more than 300 times as many units

[12] `https://github.com/tensorflow/models/tree/master/lm_1b`

as our LM; at 1.04 billion parameters it has almost 2000 times as many parameters. It is a fine-tuned language model that achieves impressive perplexity scores on common benchmarks, requires a massive infrastructure for training, and pushes the boundaries of what's feasible with current hardware.

We tested the Google LM with the methodology we used to test ours.[13] Due to computational resource limitations, we did not evaluate it on the entire test set, but sampled a random selection of 500 sentences for each count of attractors (testing a single sentence under the Google LM takes around 5 seconds on average). The results are presented in Figure 4c, where they are compared to the performance of the supervised verb inflection system. Despite having an order of magnitude more parameters and significantly larger training data, the Google LM performed poorly compared to the supervised models; even a single attractor led to a sharp increase in error rate to 28.5%, almost as high as our small-scale LM (32.6% on the same sentences). While additional attractors caused milder degradation than in our LM, the performance of the Google LM on sentences with four attractors was still worse than always guessing the majority class (SINGULAR).

In summary, our experiments with the Google LM do not change our conclusions: the contrast between the poor performance of the LMs and the strong performance of the explicitly supervised objectives suggests that direct supervision has a dramatic effect on the model's ability to learn syntax-sensitive dependencies. Given that the Google LM was already trained on several hundred times more data than the number prediction system, it appears unlikely that its relatively poor performance was due to lack of training data.

6 Additional Experiments

Comparison to simple recurrent networks: How much of the success of the network is due to the LSTM cells? We repeated the number prediction experiment with a simple recurrent network (SRN) (Elman, 1990), with the same number of hidden units. The SRN's performance was inferior to the LSTM's, but the average performance for a given number of agreement attractors does not suggest a qualitative difference between the cell types: the SRN makes about twice as many errors as the LSTM across the board (Figure 4d).

Training only on difficult dependencies: Only a small proportion of the dependencies in the corpus had agreement attractors (Figure 2e). Would the network generalize better if dependencies with intervening nouns were emphasized during training? We repeated our number prediction experiment, this time training the model only on dependencies with at least one intervening noun (of any number). We doubled the proportion of training sentences to 20%, since the total size of the corpus was smaller (226K dependencies).

This training regime resulted in a 27% decrease in error rate on dependencies with exactly one attractor (from 4.1% to 3.0%). This decrease is statistically significant, and encouraging given that the total number of dependencies in training was much lower, which complicates the learning of word embeddings. Error rates mildly decreased in dependencies with more attractors as well, suggesting some generalization (Figure 4d). Surprisingly, a similar experiment using the grammaticality judgment task led to a slight *increase* in error rate. While tentative at this point, these results suggest that oversampling difficult training cases may be beneficial; a curriculum progressing from easier to harder dependencies (Elman, 1993) may provide additional gains.

7 Error Analysis

Singular vs. plural subjects: Most of the nouns in English are singular: in our corpus, the fraction of singular subjects is 68%. Agreement attraction errors in humans are much more common when the attractor is plural than when it is singular (Bock and Miller, 1991; Eberhard et al., 2005). Do our models' error rates depend on the number of the subject?

As Figure 2b shows, our LSTM number prediction model makes somewhat more agreement attraction errors with plural than with singular attractors; the difference is statistically significant, but the asymmetry is much less pronounced than in humans. Interestingly, the SRN version of the model does show a large asymmetry, especially as the count of attractors

[13] One technical exception was that we did not replace low-frequency words with their part-of-speech, since the Google LM is a large-vocabulary language model, and does not have parts-of-speech as part of its vocabulary.

increases; with four plural attractors the error rate reaches 60% (Figure 4e).

Qualitative analysis: We manually examined a sample of 200 cases in which the majority of the 20 runs of the number prediction network made the wrong prediction. There were only 8890 such dependencies (about 0.6%). Many of those were straightforward agreement attraction errors; others were difficult to interpret. We mention here three classes of errors that can motivate future experiments.

The networks often misidentified the heads of noun-noun compounds. In (17), for example, the models predict a singular verb even though the number of the subject *conservation refugees* should be determined by its head *refugees*. This suggests that the networks didn't master the structure of English noun-noun compounds.[14]

(17) Conservation **refugees live** in a world colored in shades of gray; limbo.

(18) Information technology (IT) **assets** commonly **hold** large volumes of confidential data.

Some verbs that are ambiguous with plural nouns seem to have been misanalyzed as plural nouns and consequently act as attractors. The models predicted a plural verb in the following two sentences even though neither of them has any plural nouns, possibly because of the ambiguous verbs *drives* and *lands*:

(19) The **ship** that the player drives **has** a very high speed.

(20) It was also to be used to learn if the **area** where the lander lands **is** typical of the surrounding terrain.

Other errors appear to be due to difficulty not in identifying the subject but in determining whether it is plural or singular. In Example (22), in particular, there is very little information in the left context of the subject *5 paragraphs* suggesting that the writer considers it to be singular:

(21) Rabaul-based Japanese **aircraft make** three dive-bombing attacks.

(22) The lead is also rather long; 5 **paragraphs is** pretty lengthy for a 62 kilobyte article.

The last errors point to a limitation of the number prediction task, which jointly evaluates the model's ability to identify the subject and its ability to assign the correct number to noun phrases.

8 Related Work

The majority of NLP work on neural networks evaluates them on their performance in a task such as language modeling or machine translation (Sundermeyer et al., 2012; Bahdanau et al., 2015). These evaluation setups average over many different syntactic constructions, making it difficult to isolate the network's syntactic capabilities.

Other studies have tested the capabilities of RNNs to learn simple artificial languages. Gers and Schmidhuber (2001) showed that LSTMs can learn the context-free language a^nb^n, generalizing to ns as high as 1000 even when trained only on $n \in \{1, \dots, 10\}$. Simple recurrent networks struggled with this language (Rodriguez et al., 1999; Rodriguez, 2001). These results have been recently replicated and extended by Joulin and Mikolov (2015).

Elman (1991) tested an SRN on a miniature language that simulated English relative clauses, and found that the network was only able to learn the language under highly specific circumstances (Elman, 1993), though later work has called some of his conclusions into question (Rohde and Plaut, 1999; Cartling, 2008). Frank et al. (2013) studied the acquisition of anaphora coreference by SRNs, again in a miniature language. Recently, Bowman et al. (2015) tested the ability of LSTMs to learn an artificial language based on propositional logic. As in our study, the performance of the network degraded as the complexity of the test sentences increased.

Karpathy et al. (2016) present analyses and visualization methods for character-level RNNs. Kádár et al. (2016) and Li et al. (2016) suggest visualization techniques for word-level RNNs trained to perform tasks that aren't explicitly syntactic (image captioning and sentiment analysis).

Early work that used neural networks to model

[14] The dependencies are presented as they appeared in the corpus; the predicted number was the opposite of the correct one (e.g., singular in (17), where the original is plural).

grammaticality judgments includes Allen and Seidenberg (1999) and Lawrence et al. (1996). More recently, the connection between grammaticality judgments and the probabilities assigned by a language model was explored by Clark et al. (2013) and Lau et al. (2015). Finally, arguments for evaluating NLP models on a strategically sampled set of dependency types rather than a random sample of sentences have been made in the parsing literature (Rimell et al., 2009; Nivre et al., 2010; Bender et al., 2011).

9 Discussion and Future Work

Neural network architectures are typically evaluated on random samples of naturally occurring sentences, e.g., using perplexity on held-out data in language modeling. Since the majority of natural language sentences are grammatically simple, models can achieve high overall accuracy using flawed heuristics that fail on harder cases. This makes it difficult to distinguish simple but robust sequence models from more expressive architectures (Socher, 2014; Grefenstette et al., 2015; Joulin and Mikolov, 2015). Our work suggests an alternative strategy—evaluation on naturally occurring sentences that are sampled based on their grammatical complexity—which can provide more nuanced tests of language models (Rimell et al., 2009; Bender et al., 2011).

This approach can be extended to the training stage: neural networks can be encouraged to develop more sophisticated generalizations by oversampling grammatically challenging training sentences. We took a first step in this direction when we trained the network only on dependencies with intervening nouns (Section 6). This training regime indeed improved the performance of the network; however, the improvement was quantitative rather than qualitative: there was limited generalization to dependencies that were even more difficult than those encountered in training. Further experiments are needed to establish the efficacy of this method.

A network that has acquired syntactic representations sophisticated enough to handle subject-verb agreement is likely to show improved performance on other structure-sensitive dependencies, including pronoun coreference, quantifier scope and negative polarity items. As such, neural models used in NLP applications may benefit from grammatically sophisticated sentence representations developed in a multitask learning setup (Caruana, 1998), where the model is trained concurrently on the task of interest and on one of the tasks we proposed in this paper. Of course, grammatical phenomena differ from each other in many ways. The distribution of negative polarity items is highly sensitive to semantic factors (Giannakidou, 2011). Restrictions on unbounded dependencies (Ross, 1967) may require richer syntactic representations than those required for subject-verb dependencies. The extent to which the results of our study will generalize to other constructions and other languages, then, is a matter for empirical research.

Humans occasionally make agreement attraction mistakes during language production (Bock and Miller, 1991) and comprehension (Nicol et al., 1997). These errors persist in human acceptability judgments (Tanner et al., 2014), which parallel our grammaticality judgment task. Cases of grammatical agreement with the nearest rather than structurally relevant constituent have been documented in languages such as Slovenian (Marušič et al., 2007), and have even been argued to be occasionally grammatical in English (Zwicky, 2005). In future work, exploring the relationship between these cases and neural network predictions can shed light on the cognitive plausibility of those networks.

10 Conclusion

LSTMs are sequence models; they do not have built-in hierarchical representations. We have investigated how well they can learn subject-verb agreement, a phenomenon that crucially depends on hierarchical syntactic structure. When provided explicit supervision, LSTMs were able to learn to perform the verb-number agreement task in most cases, although their error rate increased on particularly difficult sentences. We conclude that LSTMs can learn to approximate structure-sensitive dependencies fairly well given explicit supervision, but more expressive architectures may be necessary to eliminate errors altogether. Finally, our results provide evidence that the language modeling objective is not by itself sufficient for learning structure-sensitive dependencies, and suggest that a joint training objective can be used to supplement language models on tasks for which syntax-sensitive dependencies are important.

Acknowledgments

We thank Marco Baroni, Grzegorz Chrupała, Alexander Clark, Sol Lago, Paul Smolensky, Benjamin Spector and Roberto Zamparelli for comments and discussion. This research was supported by the European Research Council (grant ERC-2011-AdG 295810 BOOTPHON), the Agence Nationale pour la Recherche (grants ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC) and the Israeli Science Foundation (grant number 1555/15).

References

Joseph Allen and Mark S. Seidenberg. 1999. The emergence of grammaticality in connectionist networks. In Brian MacWhinney, editor, *Emergentist approaches to language: Proceedings of the 28th Carnegie symposium on cognition*, pages 115–151. Mahwah, NJ: Erlbaum.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference for Learning Representations*.

Emily M. Bender, Dan Flickinger, Stephan Oepen, and Yi Zhang. 2011. Parser evaluation over local and non-local deep dependencies in a large corpus. In *Proceedings of EMNLP*, pages 397–408.

Kathryn Bock and Carol A. Miller. 1991. Broken agreement. *Cognitive Psychology*, 23(1):45–93.

Melissa Bowerman. 1988. The "no negative evidence" problem: How do children avoid constructing an overly general grammar? In John A. Hawkins, editor, *Explaining language universals*, pages 73–101. Oxford: Basil Blackwell.

Samuel R. Bowman, Christopher D. Manning, and Christopher Potts. 2015. Tree-structured composition in neural networks without tree-structured architectures. In *Proceedings of the NIPS Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*.

Bo Cartling. 2008. On the implicit acquisition of a context-free grammar by a simple recurrent neural network. *Neurocomputing*, 71(7):1527–1537.

Rich Caruana. 1998. Multitask learning. In Sebastian Thrun and Lorien Pratt, editors, *Learning to learn*, pages 95–133. Boston: Kluwer.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of EMNLP*, pages 1724–1734.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT press.

Alexander Clark, Gianluca Giorgolo, and Shalom Lappin. 2013. Statistical representation of grammaticality judgements: The limits of n-gram models. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 28–36.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and A. Noah Smith. 2016. Recurrent neural network grammars. In *Proceedings of NAACL/HLT*, pages 199–209.

Kathleen M. Eberhard, J. Cooper Cutting, and Kathryn Bock. 2005. Making syntax of sense: Number agreement in sentence production. *Psychological Review*, 112(3):531–559.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Jeffrey L. Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7(2-3):195–225.

Jeffrey L. Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Martin B. H. Everaert, Marinus A. C. Huybregts, Noam Chomsky, Robert C. Berwick, and Johan J. Bolhuis. 2015. Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, 19(12):729–743.

Robert Frank, Donald Mathis, and William Badecker. 2013. The acquisition of anaphora by simple recurrent networks. *Language Acquisition*, 20(3):181–227.

Felix Gers and Jürgen Schmidhuber. 2001. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks*, 12(6):1333–1340.

Anastasia Giannakidou. 2011. Negative and positive polarity items: Variation, licensing, and compositionality. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: An international handbook of natural language meaning*. Berlin: Mouton de Gruyter.

Yoav Goldberg and Joakim Nivre. 2012. A dynamic oracle for arc-eager dependency parsing. In *Proceedings of COLING 2012*, pages 959–976.

Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. 2015. Learning to transduce with unbounded memory. In *Advances in Neural Information Processing Systems*, pages 1828–1836.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press, Cambridge.

Armand Joulin and Tomas Mikolov. 2015. Inferring algorithmic patterns with stack-augmented recurrent nets. In *Advances in Neural Information Processing Systems*, pages 190–198.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2016. Representation of linguistic form and function in recurrent neural networks. *arXiv preprint arXiv:1602.08952*.

Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2016. Visualizing and understanding recurrent networks. In *ICLR Workshop*.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*.

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional lstm feature representations. *Transactions of the Association of Computational Linguistics*, 4:313–327.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2015. Unsupervised prediction of acceptability judgements. In *Proceedings of ACL/IJCNLP*, pages 1618–1628.

Steve Lawrence, Lee C. Giles, and Santliway Fong. 1996. Can recurrent neural networks learn natural language grammars? In *IEEE International Conference on Neural Networks*, volume 4, pages 1853–1858.

Willem J. M. Levelt, Ardi Roelofs, and Antje S. Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1):1–75.

Jiwei Li, Xinlei Chen, Eduard H. Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of NAACL-HLT 2016*, pages 681–691.

Franc Marušič, Andrew Nevins, and Amanda Saksida. 2007. Last-conjunct agreement in Slovenian. In *Annual Workshop on Formal Approaches to Slavic Linguistics*, pages 210–227.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048.

Janet L. Nicol, Kenneth I. Forster, and Csaba Veres. 1997. Subject–verb agreement processes in comprehension. *Journal of Memory and Language*, 36(4):569–587.

Joakim Nivre, Laura Rimell, Ryan McDonald, and Carlos Gomez-Rodriguez. 2010. Evaluation of dependency parsers on unbounded dependencies. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 833–841. Association for Computational Linguistics.

Laura Rimell, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of EMNLP*, pages 813–821.

Paul Rodriguez, Janet Wiles, and Jeffrey L. Elman. 1999. A recurrent neural network that learns to count. *Connection Science*, 11(1):5–40.

Paul Rodriguez. 2001. Simple recurrent networks learn context-free and context-sensitive languages by counting. *Neural Computation*, 13(9):2093–2118.

Douglas L. T. Rohde and David C. Plaut. 1999. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72(1):67–109.

John Robert Ross. 1967. *Constraints on variables in syntax*. Ph.D. thesis, MIT.

Carson T. Schütze. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago, IL: University of Chicago Press.

Richard Socher. 2014. *Recursive Deep Learning for Natural Language Processing and Computer Vision*. Ph.D. thesis, Stanford University.

Adrian Staub. 2009. On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language*, 60(2):308–327.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. In *INTERSPEECH*.

Darren Tanner, Janet Nicol, and Laurel Brehm. 2014. The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language*, 76:195–215.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. In *Advances in Neural Information Processing Systems*, pages 2755–2763.

Arnold Zwicky. 2005. Agreement with nearest always bad? `http://itre.cis.upenn.edu/~myl/languagelog/archives/001846.html`.

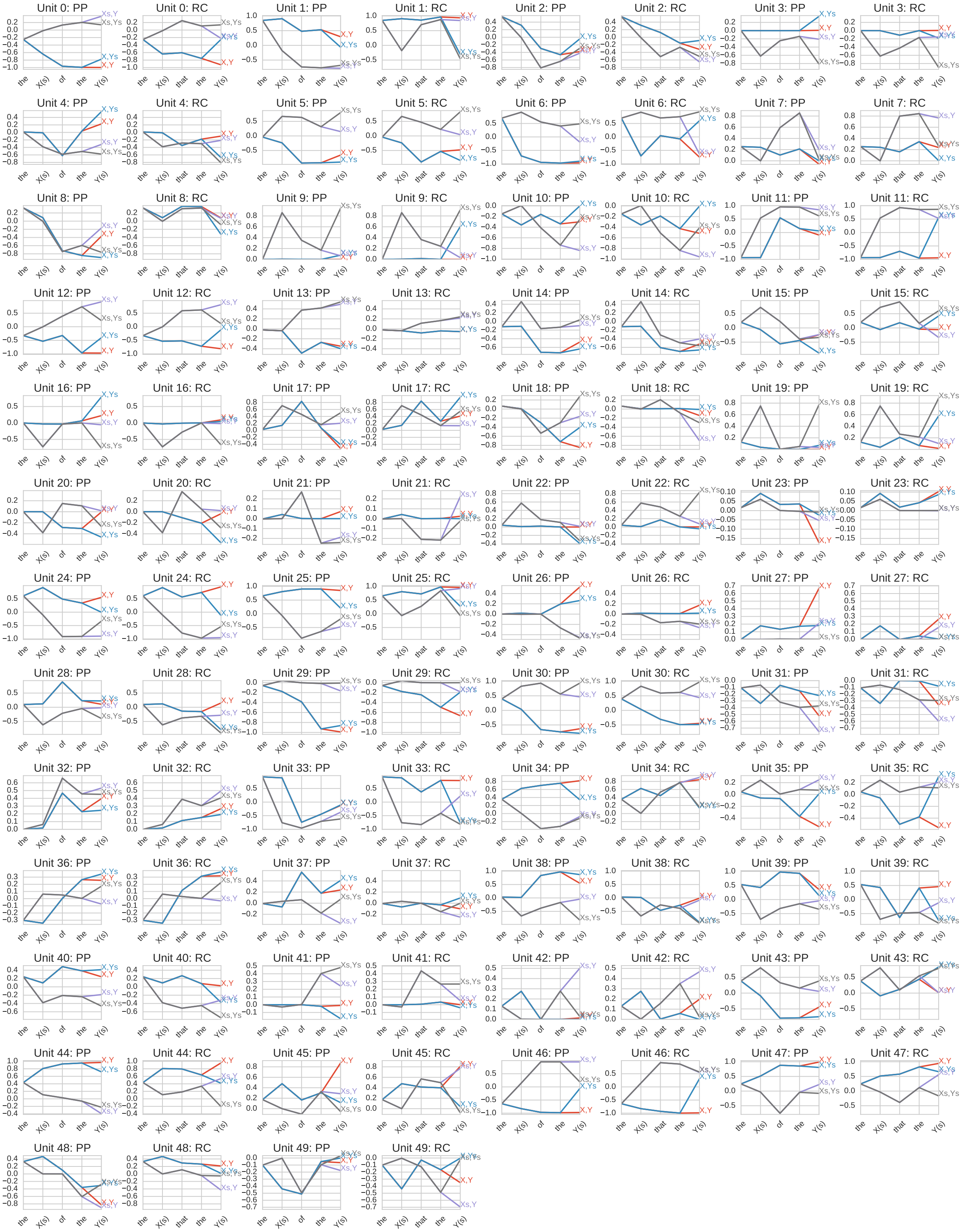
Unit 0: PP
Unit 0: RC
Unit 1: PP
Unit 1: RC
Unit 2: PP
Unit 2: RC
Unit 3: PP
Unit 3: RC
Unit 4: PP
Unit 4: RC
Unit 5: PP
Unit 5: RC
Unit 6: PP
Unit 6: RC
Unit 7: PP
Unit 7: RC
Unit 8: PP
Unit 8: RC
Unit 9: PP
Unit 9: RC
Unit 10: PP
Unit 10: RC
Unit 11: PP
Unit 11: RC
Unit 12: PP
Unit 12: RC
Unit 13: PP
Unit 13: RC
Unit 14: PP
Unit 14: RC
Unit 15: PP
Unit 15: RC
Unit 16: PP
Unit 16: RC
Unit 17: PP
Unit 17: RC
Unit 18: PP
Unit 18: RC
Unit 19: PP
Unit 19: RC
Unit 20: PP
Unit 20: RC
Unit 21: PP
Unit 21: RC
Unit 22: PP
Unit 22: RC
Unit 23: PP
Unit 23: RC
Unit 24: PP
Unit 24: RC
Unit 25: PP
Unit 25: RC
Unit 26: PP
Unit 26: RC
Unit 27: PP
Unit 27: RC
Unit 28: PP
Unit 28: RC
Unit 29: PP
Unit 29: RC
Unit 30: PP
Unit 30: RC
Unit 31: PP
Unit 31: RC
Unit 32: PP
Unit 32: RC
Unit 33: PP
Unit 33: RC
Unit 34: PP
Unit 34: RC
Unit 35: PP
Unit 35: RC
Unit 36: PP
Unit 36: RC
Unit 37: PP
Unit 37: RC
Unit 38: PP
Unit 38: RC
Unit 39: PP
Unit 39: RC
Unit 40: PP
Unit 40: RC
Unit 41: PP
Unit 41: RC
Unit 42: PP
Unit 42: RC
Unit 43: PP
Unit 43: RC
Unit 44: PP
Unit 44: RC
Unit 45: PP
Unit 45: RC
Unit 46: PP
Unit 46: RC
Unit 47: PP
Unit 47: RC
Unit 48: PP
Unit 48: RC
Unit 49: PP
Unit 49: RC

Figure 5: Activation plots for all units (see Figure 3a).