Deep Learning of Inflection and the Cell-Filling Problem

Franco Alberto Cardillo*
ILC-CNR

Marcello Ferro*
ILC-CNR

Claudia Marzi*
ILC-CNR

Vito Pirrelli*
ILC-CNR

Machine learning offers two basic strategies for morphology induction: lexical segmentation and surface word relation. The first approach assumes that words can be segmented into morphemes. Inferring a novel inflected form requires identification of morphemic constituents and a strategy for their recombination. The second approach dispenses with segmentation: lexical representations form part of a network of associatively related inflected forms. Production of a novel form consists in filling in one empty node in the network. Here, we present the results of a task of word inflection by a recurrent LSTM network that learns to fill in paradigm cells of incomplete verb paradigms. Although the task does not require morpheme segmentation, we show that accuracy in carrying out the inflection task is a function of the model's sensitivity to paradigm distribution and morphological structure.

1. Introduction

Following a morpheme-based tradition in morphological inquiry, the process of morphology induction can be defined as the task of singling out morphological formatives from fully inflected word forms. Formatives are understood to be part of the morphological lexicon, where they are accessed for word recognition, and retrieved and spelled out for word production. The view requires that a word form be segmented into meaningful sublexical signs, called *morphemes*, each contributing a separable piece of morpho-lexical content. In inflecting languages, typically this holds for regularly inflected forms, as with Italian *cred-ut-o* 'believed' (past participle, from CREDERE 'believe'), where *cred-* conveys the lexical meaning of CREDERE, and *-ut-o* is associated with morpho-syntactic features. A further assumption is that there always exists an underlying *base form* from which all other forms are spelled out. In an irregular verb form like Italian *appes-o* 'hung' (from APPENDERE), however, it soon becomes difficult to separate morpho-lexical information (the verb stem) from morpho-syntactic information (e.g. past participle). Various technical attempts have been made to circumvent these problems and restore the disrupted biuniqueness between forms and morpho-sysntactic content in non trivial inflectional systems. In fact, for too many languages, morpheme segmentation is an ill-defined task, due to notorious problems with the classical, sign-based notion of morpheme, and the non-segmental processes of introflexive (i.e. root and pattern), tonal and apophony-based morphologies. So, the assumption that

* Istituto di Linguistica Computazionale "A. Zampolli" - via G. Moruzzi 1, 56124 Pisa, Italy.
E-mail: name.surname@ilc.cnr.it

any word form can uniquely and consistently be segmented into morphemic sublexical constituents is at best dubious, and cannot be entertained as a general bootstrapping hypothesis for morphology learning.

A different formulation of the same task assumes that the lexicon consists of fully-inflected word forms and that morphology induction is the result of discovering implicative relations between them. Unknown forms are inferred through redundant analogy-based patterns between known forms, along the lines of an analogical proportion such as:

rendere 'make' :: *reso* 'made' = *appendere* 'hang' :: *appeso* 'hung'.

Support to this view comes from developmental psychology, where words are understood as the foundational elements of language acquisition, from which early grammar rules emerge epiphenomenally (Tomasello 2000; Goldberg 2003). After all, children are exposed to fully inflected forms in acquisition, and have no privileged access to underlying base forms. Besides, they appear to be extremely sensitive to sub-regularities holding between inflectionally-related forms (Bittner, Dressler, and Kilani-Schoch 2003; Colombo et al. 2004; Dąbrowska 2004; Orsolini and Marslen-Wilson 1997; Orsolini, Fanari, and Bowles 1998). Further support is lent by neurobiologically inspired computer models of language, blurring the traditional dichotomy between processing and storage (Elman 2009; Marzi et al. 2016).

In the present paper, we will mainly be concerned with issues of cognitive plausibility and inter-linguistic coverage for computational models of word generation. Our main emphasis here is not on the most effective machine learning strategy for a specific language classification task on a specific language, but on the general algorithmic requirements of the developmentally realistic task we set ourselves to (namely the 'cell-filling problem', see infra). In particular, we will focus on how these requirements are met by one of the most advanced and sophisticated models of recurrent neural networks to date, so-called Long Short Term Memory (LSTM) networks (Bengio, Simard, and Frasconi 1994; Hochreiter and Schmidhuber 1997; Malouf 2017). Hence, comparison of the system performance with other competing systems will be carried out only to the extent needed to show that the proposed architecture compares reasonably well with state of the art algorithms and to focus on its learning bias. As we are not interested in proving that our LSTM architecture performs better than other systems, but only in assessing its potential with the sparsest possible set of language-specific assumptions, issues of task-driven, language-driven and parameter-driven optimization are not addressed.

2. The cell-filling problem

To understand how word inflection can be conceptualised as a word relation task, it is useful to think of this task as a *cell-filling problem* (Ackerman and Malouf 2013; Ackerman, Blevins, and Malouf 2009). Inflected forms are traditionally arranged in so-called *paradigms*. The full paradigm of CREDERE 'believe' is a labelled set of all its inflected forms: *credere, credendo, creduto, credo* etc. In most cases, these forms take one and only one *cell*, defined as a specific combination of tense, mood, person and number features: e.g. *crede*, PRES IND, 3S. In all languages, words happen to follow a Zipfian distribution, with very few high-frequency words, and a very long tail of exceedingly rare words (Blevins, Milin, and Ramscar 2017). As a result, even high-frequency paradigms happen to be attested partially, and speakers must then be able to generalise incomplete paradigmatic knowledge. This amounts to a cell-filling problem:

given a set of attested forms in a paradigm, the speaker has to guess what other forms can fill in empty cells in the same paradigm.

2.1 Computational modelling

Borrowing Blevins' (2006) terminology, we can make a distinction between "constructive" and "abstractive" algorithms for word learning. Constructive algorithms assume that classificatory information is morpheme-based. Word forms are segmented into morphemes for training, and a classifier must learn to apply morpheme segmentation to novel forms after training. An abstractive learning algorithm, on the other hand, sees morphological structure as emerging from full forms, be they annotated with classificatory information (supervised mode) or not (unsupervised mode). From this perspective, training data consist of unsegmented word forms (strings of either letters or sounds), possibly coupled with their lexical and morpho-syntactic content. Accordingly, morphological acquisition boils down to learning from lexical representations in training, to generalise them to unknown forms. In this process, word-internal constituents can possibly emerge, either as a result of the formal redundancy of raw input data (unsupervised mode), or as a by-product of form-content mappings (supervised mode). Only abstractive machine learning models of inflection can address the cell-filling problem. Thus, we will hereafter focus on abstractive models.

A further important qualification to be made in this connection concerns the set of a-priori assumptions about the target inflection system that some abstractive algorithms avail themselves of. For example, knowledge that the target language morphology is concatenative can considerably constrain the hypothesis search space of the algorithm, which is biased to look for stem-ending patterns only. This bias has important implications for word acquisition. Although no explicit morpheme segmentation is provided in training, the way word forms are tentatively split into internal constituents brings to bear detailed information about boundary relations between constituents (Goldsmith 2001). Other a-priori biases may consist in (i) using fixed-length positional templates (Keuleers and Daelemans 2007; Plunkett and Juola 1999), or (ii) tying individual symbols (letters or sounds) to specific positions in the input representation (so-called "conjunctive" coding) (Coltheart et al. 2001; Harm and Seidenberg 1999; McClelland and Rumelhart 1981; Perry, Ziegler, and Zorzi 2007; Plaut et al. 1996), or (iii) resorting to some language-specific alignment algorithms (Albright 2002) or head-and-tail splitting procedures (Pirrelli and Yvon 1999).

We contend that a bootstrapping algorithm for morphology induction should be valued for its ability to converge on the acquisition of an inflection system with the sparsest possible set of a-priori assumptions about the underlying structure of the system, rather than for its learning bias. The ability to recognise position-independent patterns in symbolic time series, like the word *book* in *handbook*, or the verb root *mach* in German *gemacht* ('made' past participle), lies at the heart of human learning of inflection. A more human-like algorithm for morphological bootstrapping should have the capacity to adapt itself to the morphological structure of the target language. This is all the more important, if we consider that the way morpho-syntactic features are contextually realised through processes of word inflection arguably represents the widest dimension of crosslinguistic grammatical variation (somewhat belittling universal invariances along other dimensions (Evans and Levinson 2009)). Although many comprehensive catalogues of the morphological markers and patterns in a given language or languages are available (Bickel and Nichols 2005; McWorther 2001; Shosted 2006), there exists no close inventory of parametric cross-linguistic variation for inflection.

In the machine learning literature, a more principled approach to morphology induction has been taken by so-called "features and classes" approaches (McNamee, Nicholas, and Mayfield 2009; Pauw and Wagacha 2007), where a word form is represented as a set of redundantly specified n-grams, i.e. possibly overlapping substrings of *n* characters making up the input string: for example, 'wa', 'al', and 'lk' for the string *walk*. N-grams have no internal structure and may be order-independent. The algorithm may start with the hypothesis that each word form is in a class of its own, and uses a stochastic classifier to calculate the conditional probability of having a certain class (a word form) given the set of distributed n-grams associated with the class. N-grams that occur in many words will be poorly discriminative, whereas features that happen to be repeatedly associated with a few word forms only will be given a morphologically meaningful interpretation. Features and classes approaches are in a position to address the bootstrapping issue of converging on the appropriate morphological classification of input data with no a-priori learning biases. However, due to their n-gram bases representations of input forms, it is not clear how they can be used for generating a fully spelled out form from its lexical and morpho-syntactic features, as required by the cell-filling problem.

A few more recent connectionist models have addressed the problem of learning inflectional paradigms. Goldsmith and O'Brien's (2006) network takes as input a lexeme identifier and a paradigm cell from the Spanish conjugation, to output the correct realisation for the corresponding form. However, the output is not a full form, but an identifier for one of a predefined set of possible realisations: e.g. *-amos* for CANTAR 'sing', PRES IND, 1P. A more psycholinguistically inspired connectionist network is described by Thymï¿œ et al. (1994), which nonetheless applies to a few noun lemmas only, and is mostly intended to simulate human errors. Finally, Temporal Self-Organising Maps (TSOMs) have recently been proposed as models of dynamic memories for symbolic time-series. In TSOMs, words are represented as chains of specialised processing nodes, that selectively fire when specific symbols are input in specific temporal contexts. Node specialisation is the outcome of the interplay of two training principles, based on entrenchment and competition. Marzi and colleagues (2018) discuss some important properties of TSOMs trained on 6 inflection systems of different complexity (including one typologically different one). However plausible as models of abstractive paradigm-based learning of inflection morphology, TSOMs are, however, not readily amenable to being used for the cell-filling problem.

Here, we consider another different connectionist architecture, based on Long Short Term Memories (LSTMs), recently proposed by Malouf (2016, 2017) to address the cell-filling problem.

3. The Experiment

Cell-filling can be simulated by training a learning model on a number of partial paradigms, to then complete them by generating missing forms. Training consists of <lemma_paradigm cell, inflected form> pairs. A lemma is not a form (e.g. *credere* 'to believe'), but a symbolic proxy of its lexical content (CREDERE). Word inflection consists of producing a fully inflected form given a known lemma and an empty paradigm cell. It is important to appreciate that for a model to be able to produce an inflected form on the basis of a cluster of lexical and morpho-syntactic features, it has to learn the symbol sequence making up the stem of the verb in question in the first place and to combine it with the paradigmatically appropriate inflectional ending.

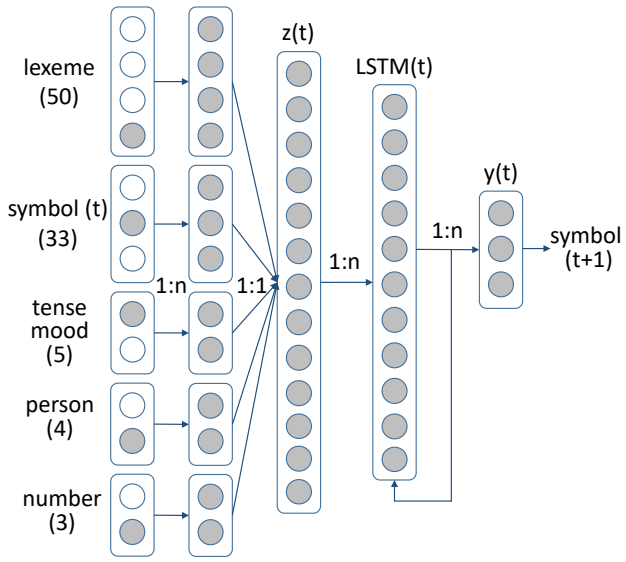

Figure 1
The network architecture. The input vector dimension is shown in brackets. Trainable dense projection matrices are shown as $1:n$, and concatenation as $1:1$.

3.1 Methods and materials

Following Malouf (2017), the LSTM network in Figure 1 is designed to take as input a lemma (e.g. CREDERE), a set of morpho-syntactic features (e.g. PRES_IND, 3, S) and a sequence of symbols (<*crede*>)[1] one symbol s_t at a time, to output a probability distribution over the upcoming symbol s_{t+1} in the sequence: $p(s_{t+1}|s_t$,CREDERE, PRES_IND, 3, S). To produce the form <*crede*>, we take the start symbol '<' as s_1, use s_1 to predict s_2, then use the predicted symbol to predict s_3 and so on, until '>' is predicted. Input symbols are encoded as mutually orthogonal one-hot vectors with as many dimensions as the overall number of different symbols used to encode all forms in the dataset.

The architecture was implemented in the Python language using two software libraries: Keras[2] and TensorFlow[3]. Keras is a high-level Python library that allows to define artificial neural networks using simple (even functional) APIs. Once defined the neural architecture, Keras relies on other libraries for all the numerical computations. In our case, we used the library TensorFlow (and its Python API) configured to run on the GPU. [4]

Unlike in Malouf's architecture, where morpho-syntactic features are holistically encoded in the input layer as one-hot vectors, each representing an orthogonal bundle of tense, mood, person and number features, here the morpho-syntactic features of tense, person and number are given independent one-hot vectors, whose dimensions equal the

1 '<' and '>' are respectively the start-of-word and the end-of-word symbols
2 `https://keras.io/`
3 `https://www.tensorflow.org/`
4 When running on a nVidia GeForce GTX970 with 1664 CUDA cores, a single training iteration (one folder in the leave-one-out cross-validation, with mini-batches of size 32) lasts between 15 and 30 seconds, depending on the network size and on the training language (more precisely, on the average length of the input sequences of symbols).

Table 1
Language datasets. Form length is measured by the number of orthographic symbols. In the Italian sample, the orthographic accent is encoded as a separate character (e.g. *ï¿œ* = e'). Differences between form types and cardinality of the training set are due to syncretism (particularly extensive in English).

language	*min/max form length*	*regular/irregular paradigms*	*form types/ training size*
English	2/11	20/30	208/750
German	3/11	16/34	504/750
Italian	2/12	23/27	748/750
Spanish	2/15	23/27	715/750

number of different values each feature can take. An extra dimension is added when a feature can be left uninstatiated in particular forms, as is the case with person and number features in the infinitive. No information is given about conjugation class for those languages (like Italian and Spanish) with more than one such class. This choice is motivated by the need to keep our network architecture as language-independent as possible, thus minimising recourse to those representational "tricks" that presuppose some knowledge of the language being learned. Language morphologies may differ considerably in the way morpho-syntactic feature bundles (e.g. <PRES_IND, 3, S>) are mapped onto surface markers. Some (more fusional) languages appear to realise a whole bundle of features with a single marker, whereas more agglutinative languages assign different markers to different features in the same bundle. Accordingly, a more distributed representation of morpho-syntactic features on the input layer is the most uncommitted, language-neutral option. Preliminary tests of the network performance using either style of feature encoding (i.e. "bundled" vs. "distributed") confirmed this assumption, showing that Malouf's encoding style is linguistically more biased than a distributed encoding of morpho-syntactic features is.

All input vectors are encoded by trainable dense matrices whose outputs are concatenated into the projection layer $z(t)$, which is input, in turn, to a layer of LSTM blocks (Figure 1). The LSTM layer takes as input both the information of $z(t)$, and its own output at *t–1*. Recurrent LSTM blocks are known to be able to capture long-distance relations in time series of symbols (Bengio, Simard, and Frasconi 1994; Hochreiter and Schmidhuber 1997; Jozefowicz, Zaremba, and Sutskever 2015), avoiding classical problems with training gradients of Simple Recurrent Networks (Jordan 1986; Elman 1990).

We tested our model on four comparable sets of English, German, Italian and Spanish inflected verb forms (Table 1), where paradigms are selected by sampling the highest-frequency fifty paradigms in large monolingual reference corpora (the Celex database for German and English (Baayen, Piepenbrock, and Gulikers 1995), the Italian Paisï¿œ corpus (Lyding et al. 2014), the European Spanish Subcorpus of the Spanish TenTen Corpus (www.sketchengine.co.uk)). For all languages, a fixed set of cells was chosen from each paradigm: all present indicative forms (1SIE, 2SIE, 3SIE, 1PIE, 2PIE, 3PIE), all past tense forms (1SIA, 2SIA, 3SIA, 1PIA, 2PIA, 3PIA), infinitive (i), past participle (pA), German and English present participle/Italian and Spanish gerund (pE). Each training form was administered once per epoch, and training was stopped when the training accuracy did not improve by more than a fixed threshold (results in this paper have

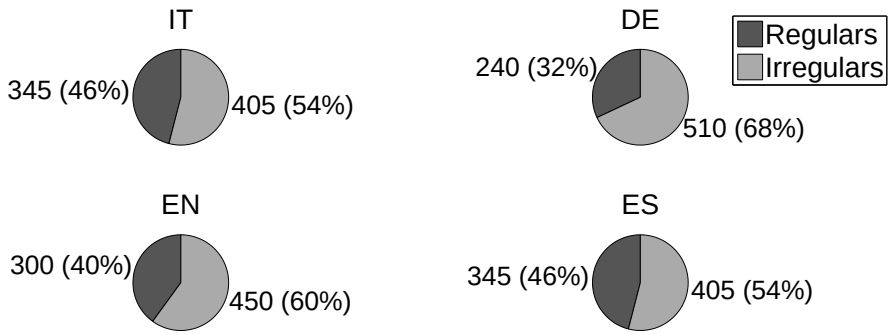

Figure 2
Composition of the four datasets (IT: Italian, DE: German, EN: English, ES: Spanish) in terms of percentage of forms belonging to regular and irregular paradigms. As expected, for all languages, the majority of top-frequency paradigms are irregular.

been obtained using the value 0.005) for a number consecutive "patience" epochs (set to five in these experiments). Although a uniform distribution is admittedly not realistic, it increases the entropy of the cell-filling problem, to define some sort of upper bound on the complexity of the task.

The four sets exhibit extensive stem allomorphy and a rich set of affixations, including circumfixation (German *ge-mach-t* 'made', past participle). Most importantly, the distribution of stem allomorphs is accountable in terms of equivalence classes of cells, forming morphologically heterogenous, phonologically poorly predictable, but fairly stable sub-paradigms (Pirrelli 2000). Selection of the contextually appropriate stem allomorph for a given cell thus requires knowledge of the form of the allomorph and of its distribution within the paradigm. For example, that 1S, 3S and 3P cells of the Italian PASSATO REMOTO always select the same stem (*e.g. pres-i*, *pres-e* and *pres-ero* of PRENDERE 'take') is a general property of the Italian conjugation system. Similarly, if an irregular English paradigm presents two stem allomorphs (say stem_1 = *find* and stem_2 = *found*), stem_2 is selected in all past tense and past participle cells, whereas stem_1 is selected elsewhere. Finally, of the four verb systems, German, Italian and Spanish present a wide variety of inflectional endings, with the German set of endings being smaller and more systematic than the other two. Among all test languages, the English verb system is of a more isolating type, with a considerably more restricted set of inflectional endings, and a plethora of bare stem forms, i.e. inflected forms with zero affixation.

4. Results

To provide a useful benchmark for the performance of the LSTM network on the cell-filling task, we used the baseline system for Task 1 of the CoNLL-SIGMORPHON-2017 Universal Morphological Reinflection shared task.[5] The model changes the base form of a verb lemma into its inflected forms through rewrite rules of increasing specificity,

5 `https://github.com/sigmorphon/conll2017` (written by Mans Hulden).

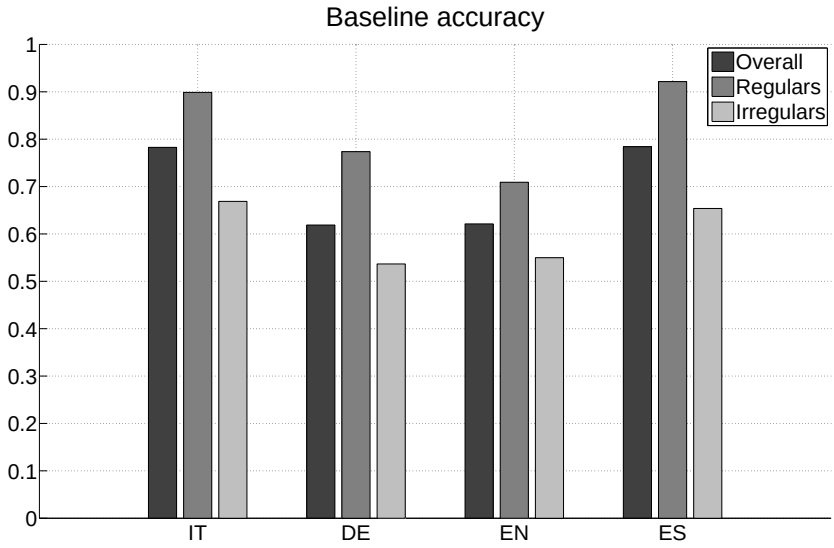

Figure 3
Per-word accuracy of the CoNLL baseline model tested on Italian (IT), German (DE), English (EN) and Spanish (ES).

automatically discovered from annotated training data. Base forms are infinitive forms for German, Italian and Spanish, and bare stems for English. To illustrate, two Italian forms such as *badare* 'to look after' and *bado* 'I look after' stand in a BASE :: PRES_IND_3S relation. The most general rule changing the former into the latter is *-are -> -o*, but more specific rewrite rules can be extracted from the same pair: *-dare -> -do*, *-adare -> -ado*, *-badare -> -bado*. The algorithm then generates the PRES_IND_3S of - say - *diradare* 'thin out', by using the most specific rewrite rule, i.e. the rewrite rule with the longest left-hand side matching *diradare* (namely *-adare > -ado*). If there is no matching rule, the base is used as a default output. It should be appreciated that the task is considerably simpler than the cell-filling problem. The algorithm can in fact infer, from the input, information about the verb base form (and, for Italian and Spanish, also about the verb conjugation class), which is to be learned from scratch in the context of the cell-filling problem.

In assessing the accuracy of the two learners, we used a leave-one-out cross-validation protocol: each form was left out of the training set, and predicted on the basis of all remaining forms. For LSTMs, this required running 750 networks, each trained on 749 exemplars and tested on the left-out exemplar.

The CoNLL baseline algorithm proves to be fairly effective for all test languages, but unevenly so (Figure 3). Somewhat unexpectedly, the model performs better on the more paradigmatically complex systems (namely Italian and Spanish, see final discussion), where regulars are inferred with remarkable accuracy (89.86% for Italian, and 91.31% for Spanish). Irregulars are predicted consistently worse (62.47% and 60.49% respectively). Accuracy on regulars is 77.92% for German and 73.67% for English, with a drop for irregulars (accuracy 51.37% and 54.89% respectively). By comparison, LSTM networks of different block size perform, on average, better than the baseline model (Figure 4). Once more, comparative accuracy varies with languages. Note that the most accurately predicted language by the baseline model (Italian) is the least accurately predicted by

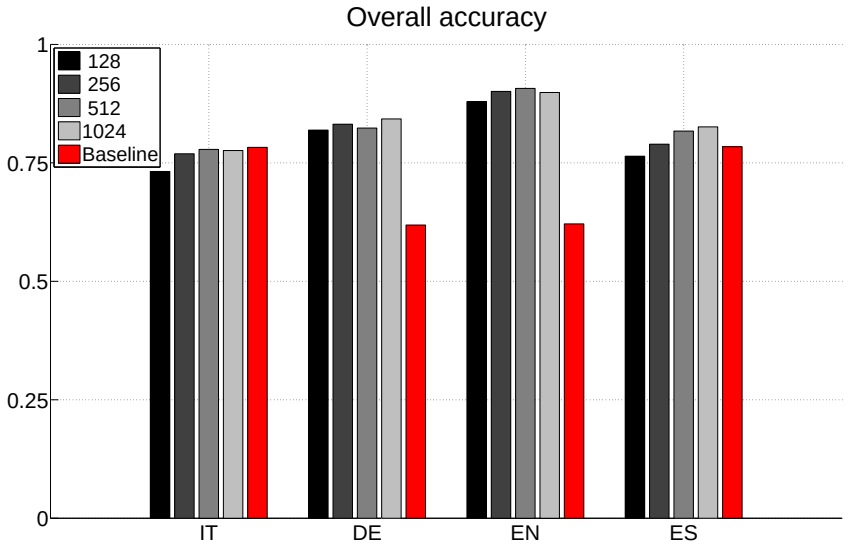

Figure 4
Per-word accuracy for Italian (IT), German (DE), English (EN) and Spanish (ES). Overall test scores are given for all LSTM network types (ordered by increasing number of LSTM blocks) and the CoNLL baseline.

LSTM networks. Nevertheless, even in this LSTM worst case, LSTM accuracy is at about the same level of the baseline accuracy. This is also true for Spanish. As for German and English, LSTM networks fare consistently better than the baseline model does (Figure 4). LSTMs average accuracy is: 73.30% on Italian, 76.77% on Spanish, 82.50% on German, and 90.20% on English forms. The CoNLL baseline accuracy is: 75.06% on Italian, 74.67% on Spanish, 59.87% on German, and 62.40% on English forms.

To check stability of our results, which could be heavily affected by LSTM initialisation point (Reimers and Gurevych 2017), we ran four different instances of the same experiment for each language, using the best performing block configuration given the language (namely 1024 for Italian, and 512 for all other languages). Accuracy scores averaged over the four istances are as follows: 74.30% on Italian (*sd* 0.4%), 78.23% on Spanish (*sd* 1%), 82.6% on German (*sd* 0.9%), and 90.93% on English (*sd* 1.3%), confirming the stability of our results.

Finally, even in case of errors at the word level, LSTMs proves to approximate the missed target more closely than the baseline does, as shown by the average edit distance between targets and outputs for all systems (Figure 5).

5. Discussion

The Cell-filling problem is a non trivial language learning task, especially in a situation where the speaker is exposed to a sample of the most frequent verb paradigms only, which include the most irregular ones in the conjugation system of all our test languages (Table 1). In a few cases, the problem has admittedly no solution. For example, in our test languages it is simply impossible to predict the first person singular of the present indicative of the auxiliary BE, on the basis of information from all remaining cells of the

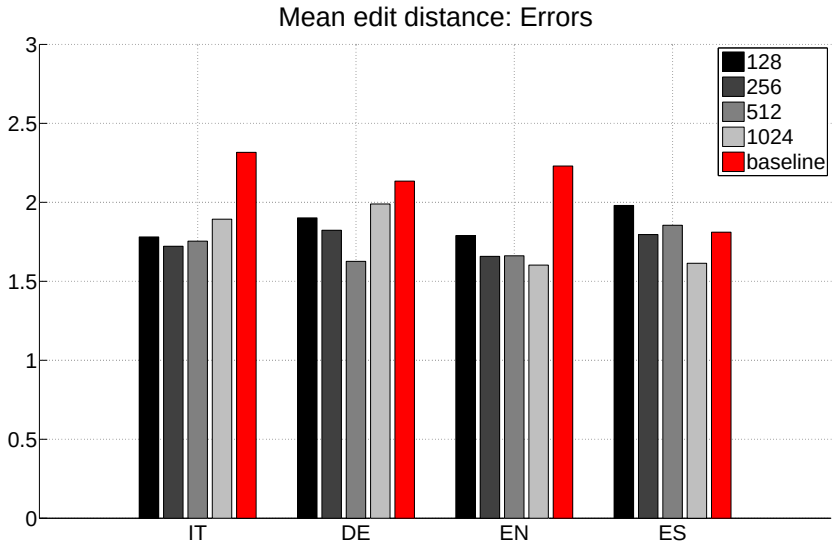

Figure 5
Mean edit distance between target form and output form for Italian (IT), German (DE), English (EN) and Spanish (ES). Distances are given for all LSTM network types (ordered by increasing number of LSTM blocks) and the CoNLL baseline.

paradigm. Another case in point is the German past participle *gefunden* 'found', which is the only form of FINDEN containing the stem alternant *fund-*, and has no other form with the same alternation pattern in our dataset. In fact, here we were not interested in replicating realistic and ecologically plausible learning conditions in child language maturation. Rather, we wanted to focus on the comparative difficulty of the task across a few languages, with the aim to assessing the ways in which LSTM networks address the logical problem of inferring novel inflected forms on the basis of the cumulative knowledge of their paradigm companions. Our experimental results clearly confirm the difficulty of the task. A powerful deep learning algorithm like an LSTM network compares well with the CoNLL baseline model, but it does not invariably outperform it. In this section, we discuss strengths and weaknesses of the two learning models.

The CoNLL baseline model is strongly reminiscent of Albright and Hayes' (2003) Minimal Generalisation Learner (MGL). MGL was, in fact, originally designed to infer Italian infinitives from first singular present indicative forms (Albright 2002). In the CoNLL model, inference goes in the opposite direction: from base forms to all other paradigm cells. Incidentally, this inferential relation is much more informative than the one adopted by Albright,[6] as the Italian infinitive contains information about the verb's thematic vowel, which is neutralised in the first singular present indicative forms, where the thematic vowel cancels out.

The CoNLL model has a clear analogy-based bias. Verb stems ending in the same way (compare - say - Italian *tend-ere* 'tend', *rend-ere* 'turn', and *prend-ere* 'take') undergo

6 In (Albright 2002), choice of the first person singular in the Italian present indicative was motivated precisely by the goal to investigate how easily the appropriate conjugation class of an Italian verb form can be predicted in those contexts where information of the thematic vowel is missing.

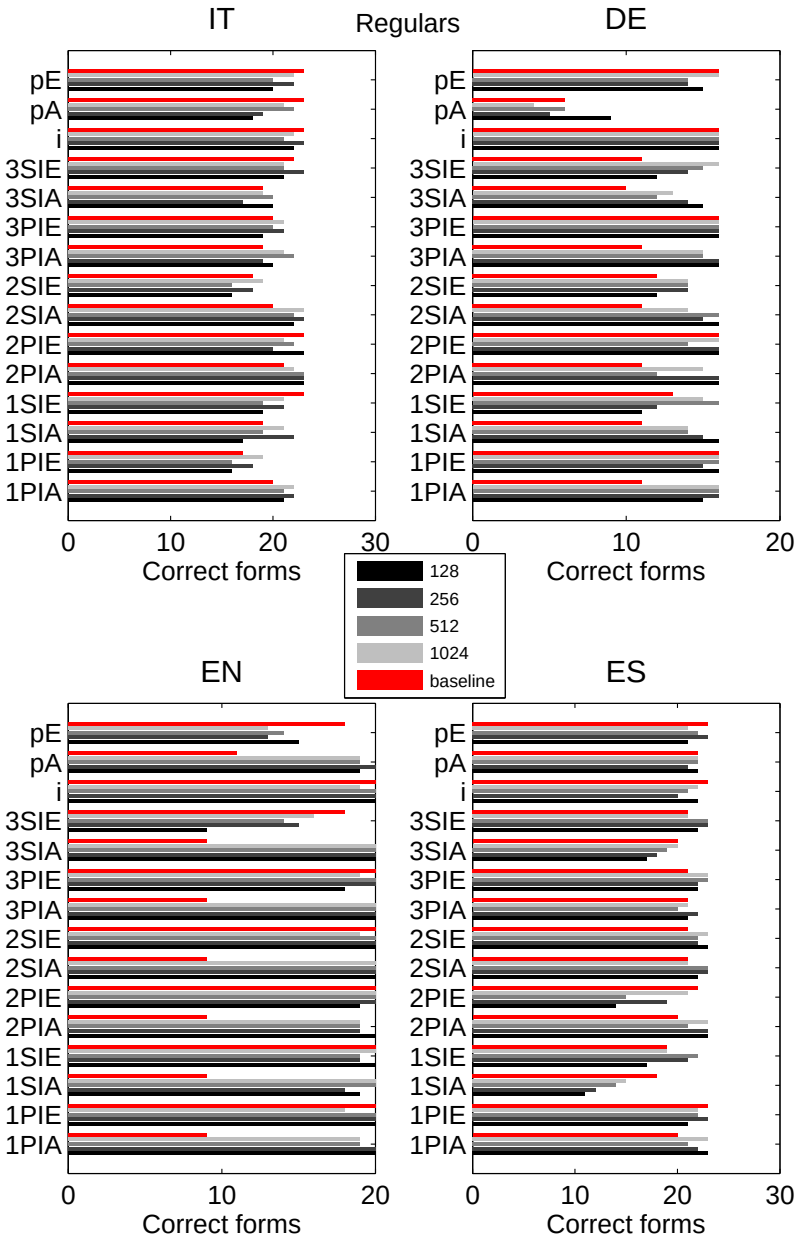

Figure 6
Per-cell accuracy of LSTMs and CoNLL baseline on regular paradigms in the four test languages.

the same allomorphic readjustment (respectively, *teso* 'tended', *reso* 'turned' and *preso* 'taken' in the past participle). In Italian, unpredictable stem allomorphy affects, in the vast majority of cases, the final part of the verb stem, and it is historically motivated by the operation of local phonological rules (Burzio 2004; Pirrelli 2000). Overall, the number of stems undergoing the same irregular stem formation processes is com-

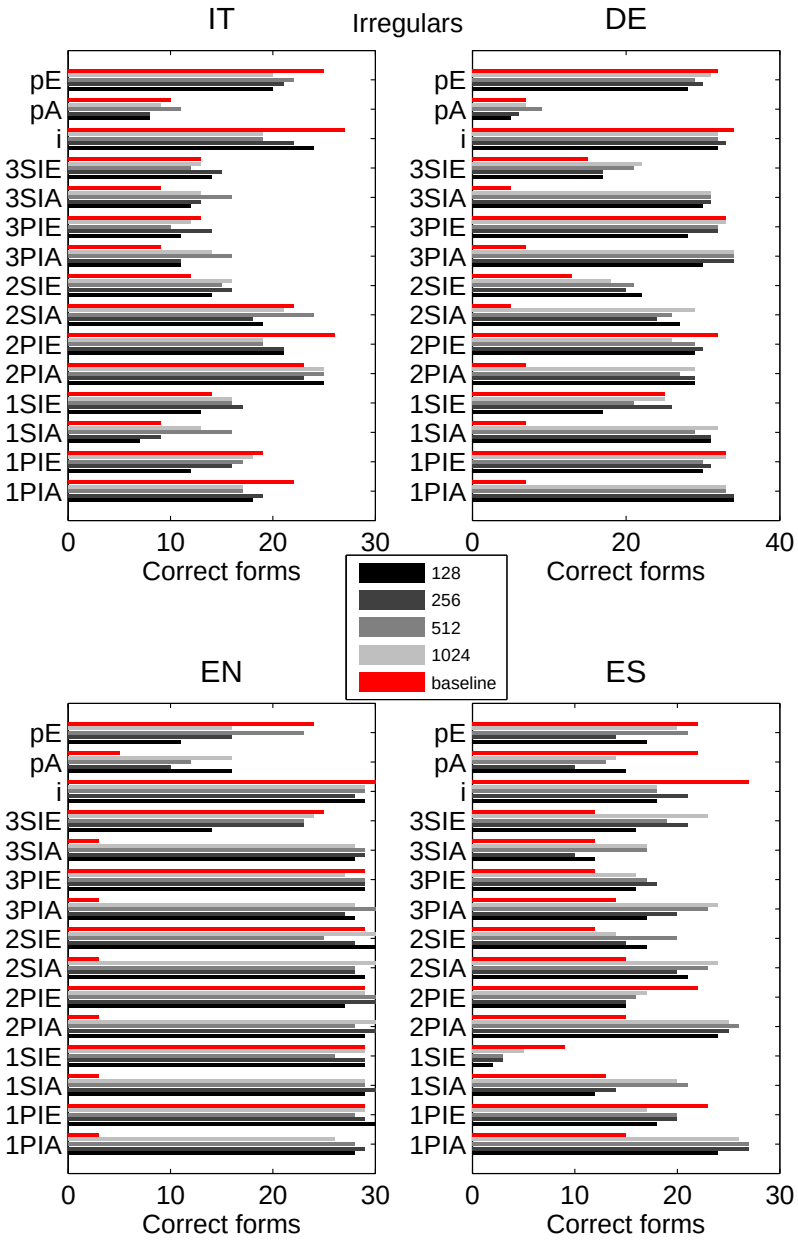

Figure 7
Per-cell accuracy of LSTMs and CoNLL baseline on irregular paradigms in the four test languages.

paratively large. Such a densely populated analogical space makes it highly likely that an unknown target form (e.g. *preso*) undergoes the same allomorphic process of the paradigmatically corresponding form of another paradigm (e.g. *reso*) whose base stem (*rend-*) ends in the same way as the target base stem (*tend-*). This makes the CoNLL baseline a powerful generalisation algorithm. Unsurprisingly, the same local

generalisation strategy is successful for Spanish too. However, in both languages, it fails to deal with stem allomorphy involving vowel apophony (as in *esco/usciamo* 'I go out/we go out'), or diphthongisation (as in Italian *vengo/vieni* 'I come/you (2S) come', and Spanish *cuento/contar* 'I count/to count', *vuelvo/volver* 'I come back/to come back'), where phonological changes are not limited to the characters immediately preceding the stem boundary.

On the other hand, it turns out that the CoNLL baseline is a much weaker learner of German and English, where stem allomorphy is not confined to the stem boundary, and is more consistently distributed throughout the paradigm. This is shown in Figures 6 (for regular paradigms) and 7 (for irregular paradigms), plotting the per cell accuracy of LSTMs and the baseline for all test languages. In all past tense and past participle cells of both German and English, where stem allomorphy applies systematically, the baseline algorithm is considerably less accurate than LSTMs are. To illustrate, let us focus on how the baseline algorithm infers the past participle of SAY. Our training set contains one maximally overlapping stem: PLAY. However, this is a "false" analogical friend to SAY, yielding the wrong past participle form **sayed*. This shows a general over-regularisation bias of the baseline, extensively witnessed by most (irregular) ablauting forms in German, which are over-regularised by the CoNLL baseline (e.g. **nemmst* for *nimmst* 'you take', *gebt* for *gibt* 'it gives').

For sure, over-regularised outcomes are plausible and frequent in child production of inflection (Clahsen, Hadler, and Weyerts 2004). A larger set of training exemplars than the one used here, including, e.g., forms of PAY (past participle *paid*), is expected to provide the evidence needed to inflect SAY correctly. Be that as it may, for our present purposes, a detailed comparison of the results of LSTMs with the CoNLL baseline is useful to understand more of the underlying morphological structure of our training data, as well as LSTMs' learning bias.

5.1 Regulars vs. irregulars

In dealing with German and English irregularly inflected forms, the purely syntagmatic approach of the CoNLL baseline, deriving all inflected forms from an underlying base, is too surface-oriented and misses some significant non local constraints. Simply put, the orthotactic/phonotactic structure of Germanic stems is less criterial for stem allomorphy than the orthotactic/phonotactic structure of Romance stems. A much more robust generalisation strategy for the two Germanic languages is to exploit the larger formal paradigmatic syncretism of their verb system.

Although LSTMs have no information about the morphological structure of input forms, they are considerably more robust than our baseline in this respect. Memory resources allowing, LSTMs appear to keep track of two types of syntagmatic constraints: short range phonological/orthotactic patterns (preventing the network to output phonotactically implausible strings such as **seemng* for target *seeming*), as well as longer range morphotactic constraints, covering the sequential structure of prefixes, stems and suffixes. Another, related issue of some theoretical interest, is to assess whether LSTMs are able to enforce global, paradigmatic constraints, whereby *all* paradigmatically-related forms contribute to fill in gaps in the same paradigm. In the end, knowledge that a paradigm contains a few stem allomorphs is good reason for a speaker to produce a stem allomorph in other (empty) cells. The more systematic the distribution of stem alternants is across the paradigm, the easier for the speaker to fill in empty cells. In this respect, German and English conjugations prove to be

paradigmatically well-behaved. Several pieces of evidence show that LSTMs are able to discover at least a few syntagmatic and paradigmatic redundancies of this kind.

First, output paradigm cells play a distinctive role in driving the generalisation bias of LSTMs. In the baseline model, a target inflected form is produced by finding the base analogue in the training data that best fits the target base. Whenever a target inflected form is produced on the basis of the best analogue to its base, the same form is output in all other cells where the best analogue happens to have identical forms. For example, if BASE::SAY best matches BASE::PLAY, SAY will be inflected as **sayed* in all part participle and past tense cells, where PLAY is inflected as *played*. This is because MGL generalisation is based on matching *input conditions* only. This is not the case for LSTMs. For example, the form **maken* is wrongly produced for PAST_PART::make, but *made* is correctly output for all past tense forms. This provides a strong indication that generalisation is also based on output cells, not on input conditions only.

A second related issue is whether LSTMs can develop global constraints on the distribution of stem allomorphs across the paradigm. We find some evidence that this is the case in analysing patterns of errors. Occasionally, past tense forms are wrongly output in past participle cells. So we find *wrote* for *written*, *took* for *taken* and *began* for *begun*. However, this pattern of errors is rather unsystematic. We suspect that it may simply be due to the repeated association of past tense forms with the feature PAST in the input vector.

What role does morphological structure play in the LSTM generalisation bias? Due to the predictive nature of the production task and the LSTM re-entrant layer of temporal connectivity, the network develops a left-to-right sensitivity to upcoming symbols, with per-symbol accuracy being a function of the network confidence about the next output symbol.[7] As we saw, this sensitivity to sequential patterns is responsible for the network's control on orthotactically/phonotactically plausible sequences. To assess the correlation between per-symbol accuracy and "perception" of morphological structure, we used Generalised Additive Models (GAM) interpolating the "average" accuracy of the learning algorithm in producing an upcoming symbol as a function of the symbol position to the inflectional boundary of each form. Results are not unequivocal, as illustrated in Figure 8.

The regression plots of Figure 8 show a clear structural effect of the distance to the stem-ending boundary of LSTM output symbols in German, Italian and Spanish, where accuracy drops to its minimum value around the morpheme boundary (corresponding to the 0 value on the x axis of the regression plots). In Italian, the most apparent difference between LSTMs (dotted lines) and the baseline (red solid line) is observed across the inflectional endings, where the per-symbol accuracy of the baseline is significantly lower than the accuracy of 256, 512 and 1028 block LSTMs. Nonetheless, per-word accuracy scores on Italian are higher in the CoNNL baseline than in LSTMs. In Spanish, the 256, 512, 1028 block LSTMs perform consistently better than the baseline. In German, the overall advantage of LSTMs over the baseline is marked by the characteristically U-shaped curve of per-symbol accuracy at the stem-ending boundary. Once more, this position marks a point of structural discontinuity in inflected verb forms. Intuitively, production of an inflected form by an LSTM network is fairly easy at the beginning of

7 For each position in the target verb form, a matching output letter is given a score of 1, and a non matching output letter a score of 0. An average score of 1 in the model means that the letter in that position is always correctly predicted, and an average score of 0 means that it is always missed.

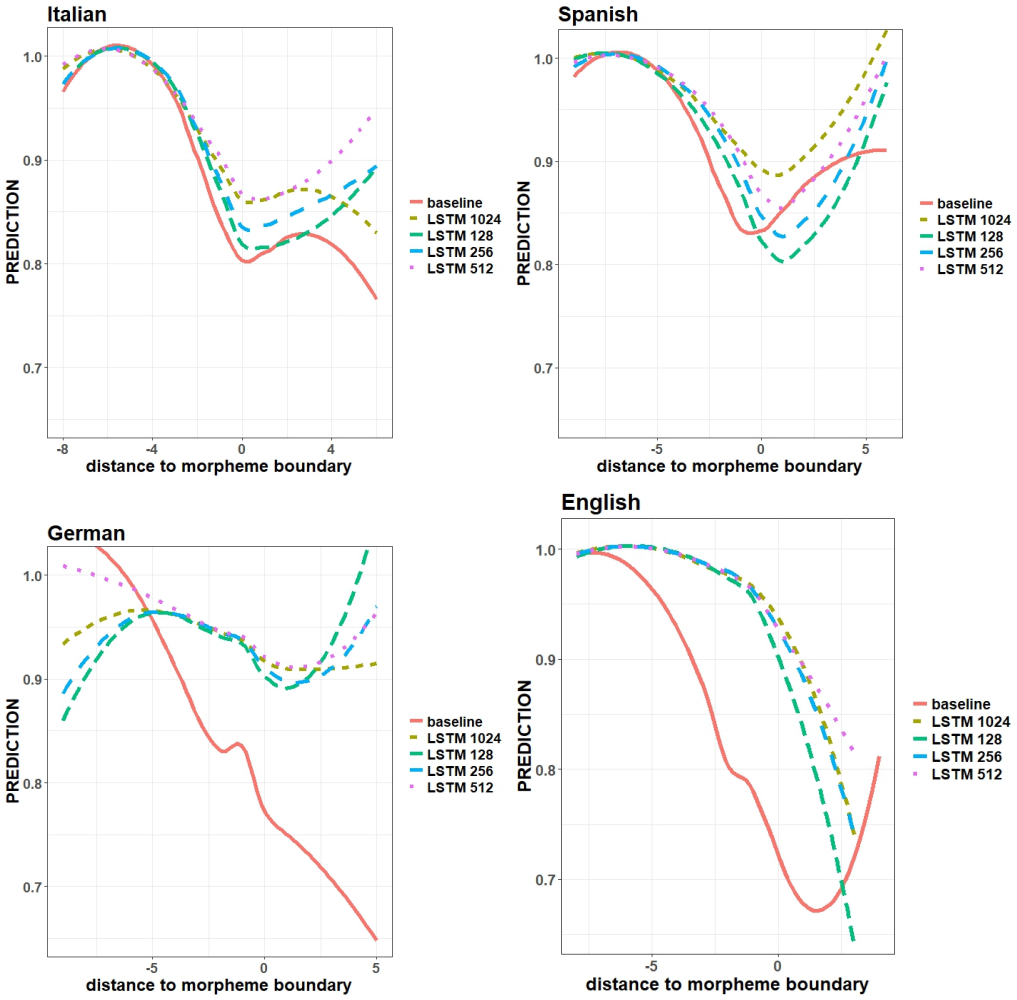

Figure 8
Regression plots of the interaction between distance to morpheme boundary (between stem and inflectional ending) and learning model in a GAM fitting per-symbol prediction accuracy in the four test languages.

the stem, but it soon gets more difficult when approaching the morpheme boundary, particularly with irregulars.

On the other hand, the English plot provides little or no evidence of structure sensitivity. The apparent U-shaped profile of the baseline does not denote a greater accuracy on long inflection endings relative to LSTM accuracy. It is merely a nonlinear interpolation effect making up for the poor performance of the baseline algorithm on English (irregular) stems compared with the corresponding endings. In fact, all LSTM models are significantly more accurate than the baseline on a long inflectional ending such as *-ing* (Figure 9, left).

The evidence has non trivial implications from a typological point of view. In a comprehensive comparison of nearly two dozen languages (in the Indo-European, Ugro-Finnic and Semitic families plus Turkish), Bittner and colleagues (2003) arrive

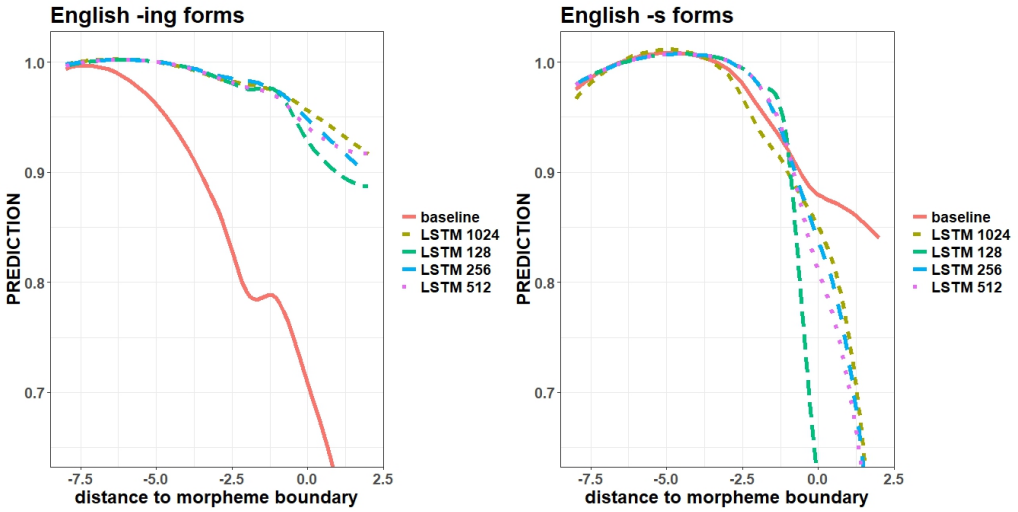

Figure 9
Regression plots of the interaction between distance to morpheme boundary (between stem and inflectional ending) and learning model in a GAM fitting per-symbol prediction accuracy in English *ing*-forms (left) and *s*-forms (right).

at the conclusion that acquisition of inflection is crucially conditioned by typological factors such as richness, uniformity and transparency of inflectional paradigms. They provide the following schema, where some European languages are arranged along a typological continuum ranging from the inflecting-fusional type (left) to the more isolating type (right):

Lithuanian→Greek→Russian→Croatian→Italian→
Spanish→German→Dutch→French→English.

The implication of this schema for our concerns is that an English inflected form provides, as such, little evidence of structural discontinuity. It is then to be expected that we find sparser evidence of processing uncertainty for symbol prediction at the morpheme boundary in a more isolating language like English. This is confirmed by evidence from child language acquisition of English inflection (Haegeman 1995; Phillips 1996), showing that English children tend to omit *s*-marking in the realisation of present indicative third singular forms, due to the overwhelming pressure of base forms in the same subparadigm. LSTMs, unlike CoNLL baseline, show a similar behaviour (Figure 9, right).[8] Note, finally, that the typological hierarchy above is somewhat mirrored by the accuracy results we obtained with LSTMs (Figure 4), where Italian is the most difficult language to be generalised over in production, and English is the easiest one. The CoNLL baseline does not seem to follow the same hierarchy. That more irregular paradigms are more difficult to learn appears to match the intuition that the morphologies of some languages are more complex than those of other languages.

8 This typological effect is somewhat amplified by the criteria we adopted for defining the position of the stem-ending boundary. For example, following Aronoff (1994), we consider an irregular past participle like English *made* as a full allomorphic stem, with no ensuing affixation.

6. Concluding remarks

The cell-filling problem addresses the ecological, developmentally motivated task of inferring novel inflected forms based on evidence of familiar forms. Other (simpler) models have been proposed in the literature to account for form-meaning mapping in Morphology (Baayen et al. 2011; Plaut and Gonnerman 2000, among others). Nevertheless, we do not know of any other artificial neural networks that can simulate word inflection as a cell-filling task. Unlike more traditional connectionist architectures (Rumelhart and McClelland 1987), recurrent LSTMs do not presuppose the existence of underlying base forms, but they learn possibly alternating stems upon exposure to linguistically annotated full forms. Admittedly, the use of orthogonal one-hot vectors for lemmas, unigram temporal series for inflected forms, and abstract morpho-syntactic features as a proxy of context-sensitive functional agreement effects, are crude representational short-hands. Nonetheless, in tackling the task, LSTMs prove to be able to orchestrate different sources of word knowledge, well beyond pure surface word relations: namely morphological structure (stem-affix boundaries), paradigm organisation and degrees of (ir-)regularity in stem formation. Acquisition of different inflectional systems may require a different balance of all these pieces of knowledge.

Unlike more *ad hoc* algorithms, LSTMs appear to be flexible and powerful enough to be able to adapt their learning strategy to the specific properties of inflectional systems of different complexity. This strikes us as an important bonus of LSTMs. In addressing a task like the cell-filling problem, which does not seem to require information about very long sequences of input symbols, LSTMs prove to be able to discover other complex constraints than just sequential or syntagmatic ones, using memory of input forms in complementary distribution as global generalisation patterns. Having said that, we should also emphasise that that the cell-filling problem turned out to be a rather recalcitrant and challenging task even for a powerful machine learning technology like LSTMs. We gathered sparse evidence that LSTMs can develop a stem variable capturing the sweepingly systematic patterns of stem distribution across paradigm cells. Like more traditional associative connectionist networks, it looks like LSTMs can learn a universally quantified one-to-one mapping relation only if this relation is illustrated with respect to each possible input/output pairs (Marcus 2001). Hence, even when an LSTM network is exposed to a number of paradigms instantiating the same pattern of stem distribution, the pattern is not readily extended to the unknown form of a partially filled in paradigm. We expect more experiments on typologically more diverse languages to be needed before the issue of the cognitive plausibility of LSTMs as models of the human word processor can be assessed on a firmer empirical basis.

References

Ackerman, Farrell, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In James P. Blevins and Juliette Blevins, editors, *Analogy in Grammar: Form and Acquisition*. Oxford University Press.

Ackerman, Farrell and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89(3):429–464.

Albright, Adam. 2002. Islands of reliability for regular morphology: Evidence from italian. *Language*, 78(4):684–709.

Albright, Adam and Bruce Hayes. 2003. Rules vs. analogy in english past tenses: A computational/experimental study. *Cognition*, 90(2):119–161.

Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes*. Number 22.

Baayen, R. Harald, Petar Milin, Dusica Filipović Đurđević, Peter Hendrix, and Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on

naive discriminative learning. *Psychological review*, 118(3):438–481.
Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers, 1995. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.
Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.
Bickel, Balthasar and Johanna Nichols. 2005. Inflectional synthesis of the verb. In David Gil Martin Haspelmath, Matthew S. Dryer and Bernard Comrie, editors, *The World Atlas of Language Structures*. Oxford University Press, pages 94–97.
Bittner, Dagmar, Wolfgang U. Dressler, and Marianne Kilani-Schoch, editors. 2003. *Development of Verb Inflection in First Language Acquisition: a cross-linguistic perspective*. Mouton de Gruyter, Berlin.
Blevins, James P. 2006. Word-based morphology. *Journal of Linguistics*, 42(3):531–573.
Blevins, James P., Petar Milin, and Michael Ramscar. 2017. The zipfian paradigm cell filling problem. In Ferenc Kiefer, James P. Blevins, and Huba Bartos, editors, *Morphological Paradigms and Functions*. Brill, Leiden.
Burzio, Luigi, 2004. *Paradigmatic and syntagmatic relations in Italian verbal inflection*, volume 258, pages 17–44. John Benjamins, Amsterdam-Philadelphia.
Clahsen, Harald, Meike Hadler, and Helga Weyerts. 2004. Speeded production of inflected words in children and adults. *Journal of child language*, 31(3):683–712.
Colombo, Lucia, Alessandro Laudanna, Maria De Martino, and Cristina Brivio. 2004. Regularity and/or consistency in the production of the past participle? *Brain and language*, 90(1):128–142.
Coltheart, Max, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. Drc: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.
Dąbrowska, Ewa. 2004. Rules or schemas? evidence from polish. *Language and cognitive processes*, 19(2):225–271.
Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.
Elman, Jeffrey L. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4):547–582.
Evans, Nicholas and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, (32):429–92.
Goldberg, Adele E. 2003. Constructions: a new theoretical approach to language. *Trends in cognitive sciences*, 7(5):219–224.
Goldsmith, John. 2001. Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2):153–198.
Goldsmith, John and Jeremy O'Brien. 2006. Learning inflectional classes. *Language Learning and Development*, 2(4):219–250.
Haegeman, Liliane. 1995. Root infinitives, tense, and truncated structures in dutch. *Language acquisition*, 4(3):205–255.
Harm, Michael W. and Mark S. Seidenberg. 1999. Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological review*, 106(3):491.
Hochreiter, Sepp and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
Jordan, Michael. 1986. Serial order: A parallel distributed processing approach. Technical Report 8604, University of California.
Jozefowicz, Rafal, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2342–2350, Lille, France, 07–09 July.
Keuleers, Emmanuel and Walter Daelemans. 2007. Memory-based learning models of inflectional morphology: A methodological case-study. *Lingue e linguaggio*, 6(2):151–174.
Lyding, Verena, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The paisá corpus of italian web texts. Proceedings of the 9th Web as Corpus Workshop (WaC-9)@ EACL 2014, pages 36–43, Gothenburg, Sweden, April, 26. Association for Computational Linguistics.
Malouf, Robert. 2016. Generating morphological paradigms with a recurrent neural network. *San Diego Linguistics Papers*, (6):122–129.
Malouf, Robert. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458.

Marcus, Gary. 2001. *The algebraic mind*. MIT Press.
Marzi, Claudia, Marcello Ferro, Franco Alberto Cardillo, and Vito Pirrelli. 2016. Effects of frequency and regularity in an integrative model of word storage and processing. *Italian Journal of Linguistics*, 28(1):79–114.
Marzi, Claudia, Marcello Ferro, Oaufae Nahli, Patrizia Belik, Stavros Bompolas, and Vito Pirrelli. 2018. Evaluating inflectional complexity crosslinguistically: a processing perspective. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, Miyazaki (Japan), 7-12 May.
McClelland, James L. and David E. Rumelhart. 1981. An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological review*, 88(5):375.
McNamee, Paul, Charles Nicholas, and James Mayfield. 2009. Addressing morphological variation in alphabetic languages. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 75–82, Boston, MA, USA, July 19 - 23. ACM.
McWorther, John. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology*, (5):125–166.
Orsolini, Margherita, Rachele Fanari, and Hugo Bowles. 1998. Acquiring regular and irregular inflection in a language with verb classes. *Language and cognitive processes*, 13(4):425–464.
Orsolini, Margherita and William Marslen-Wilson. 1997. Universals in morphological representation: Evidence from italian. *Language and Cognitive Processes*, 12(1):1–47.
Pauw, Guy De and Peter Waiganjo Wagacha. 2007. Bootstrapping morphological analysis of gĩkũyũ using unsupervised maximum entropy learning. In *Eighth Annual Conference of the International Speech Communication Association*.
Perry, Conrad, Johannes C. Ziegler, and Marco Zorzi. 2007. Nested incremental modeling in the development of computational theories: the cdp+ model of reading aloud. *Psychological review*, 114(2):273.
Phillips, Colin. 1996. Root infinitives are finite. In *Proceedings of the 20th annual Boston University conference on language development*, pages 588–599.
Pirrelli, Vito. 2000. *Paradigmi in morfologia. Un approccio interdisciplinare alla flessione verbale dell'italiano*. Istituti Editoriali e Poligrafici Internazionali, Pisa.
Pirrelli, Vito and François Yvon. 1999. The hidden dimension: a paradigmatic view of data-driven nlp. *Journal of Experimental & Theoretical Artificial Intelligence*, 11(3):391–408.
Plaut, David C. and Laura M. Gonnerman. 2000. Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, 15(4/5):445–485.
Plaut, David C., James L. McClelland, Mark S. Seidenberg, and Karalyn Patterson. 1996. Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological review*, 103(1):56.
Plunkett, Kim and Patrick Juola. 1999. A connectionist model of english past tense and plural morphology. *Cognitive Science*, 23(4):463–490.
Reimers, Nils and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.
Rumelhart, David E. and James L. McClelland. 1987. On learning the past tenses of english verbs. In David E. Rumelhart, James L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing. Explorations in the Microstructures of Cognition*, volume 2 Psychological and Biological Models. MIT Press, pages 216–271.
Shosted, Ryan. 2006. Correlating complexity: a typological approach. *Linguistic Typology*, (10):1–40.
Thymé, Ann, Farrell Ackerman, and Jeff Elman, 1994. *Finnish Nominal Inflection. Paradigmatic Patterns and Token Analogy*, volume 26, page 445. John Benjamins Publishing Company.
Tomasello, Michael. 2000. The item-based nature of children's early syntactic development. *Trends in cognitive sciences*, 4(4):156–163.