The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

Universal Recurrent Neural Network Grammar

Chinmay Choudhary
National University of Ireland
Newcastle, Galway
c.choudhary1@nuigalway.ie

Colm O'riordan
National University of Ireland
Newcastle, Galway
colm.oriordan@nuigalway.ie

Abstract

Modern approaches to Constituency Parsing are mono-lingual supervised approaches which require large amount of labelled data to be trained on, thus limiting their utility to only a handful of high-resource languages. To address this issue of data-sparsity for low-resource languages we propose **Universal Recurrent Neural Network Grammars (UniRNNG)** which is a multi-lingual variant of the popular Recurrent Neural Network Grammars (RNNG) model for constituency parsing. **UniRNNG** involves *Cross-lingual Transfer Learning* for Constituency Parsing task. The architecture of UniRNNG is inspired by *Principle and Parameter* theory proposed by Noam Chomsky. UniRNNG utilises the linguistic typology knowledge available as feature-values within WALS database, to generalize over multiple languages. Once trained on sufficiently diverse polyglot corpus **UniRNNG** can be applied to any natural language thus making it *Language-agnostic* constituency parser. Experiments reveal that our proposed **UniRNNG** outperform state-of-the-art baseline approaches for most of the target languages, for which these are tested.

Keywords: Constituency Parsing, Cross-lingual Transfer-learning

1 Introduction

Noam Chomsky proposed the hypothesis of **Universal Grammar (UG)** (Chomsky, 1986; Cook and Newson, 2014) which states that all human languages, while being superficially as diverse as they are, share some fundamental similarities. Thus he argues that deep down the specific grammars of various natural languages, there exists a *Universal Grammar*. Since then many linguists (Baker, 2008; Fodor and Sakas, 2004; Tomasello, 2005; Pinker, 1995; Fodor, 2001) attempted to outline the *principles and parameters* of this *Universal Grammar* manually, but with very limited success. If it is nearly impossible to identify and outline UG manually due to its anticipated large size and complexity (Roberts and Holmberg, 2005; Kayne, 2012; Cinque and Rizzi, 2010; Shlonsky, 2010), we can use a neural network to learn these automatically.

Recently Recurrent Neural Network based models for parsing (eg: *Recurrent Neural Network Grammars (RNNG)*(Dyer et al., 2016)) are proven to do excellent job in automatically learning and encoding (as model-parameters) the grammar of any language directly from its tree-bank corpus. This inspires us to make following assumption:

A Recurrent Neural Network based multi-lingual parser trained on a diverse polyglot treebank corpus would learn and encode the *Universal Grammar* as its model-parameters.

Based on this assumption, we propose **Universal Recurrent Neural Network Grammar (UniRNNG)** which is a multi-lingual variant of Dyer's RNNG model (Dyer et al., 2016). The architecture of **UniRNNG** is indeed inspired by the *Principle and Parameter* framework (Chomsky, 1993) advocated by linguists *Noam Chomsky* and *Howard Lasnik*. Hence unlike Dyer's RNNG, our proposed model comprises of two sets of model-parameters α and β. α would encode *Universal Principles* which are shared by all the languages and β would encode *Parameters*

which are tuned to specific language of the sentence being parsed during run-time.

Our proposed model involves *Cross-lingual Transfer Learning (CLT)* from a polyglot corpus of high-resource source-languages to a low-resource target language. CLT has extensively been applied to numerous NLP-tasks including Dependency Parsing (Daniel et al., 2017a; Zeman et al., 2018a), Natural Language Inference (Conneau et al., 2018; Singh et al., 2019; Huang et al., 2019; Doval et al., 2019), Question Answering (Liu et al., 2019; Lee and Lee, 2019; Lewis et al., 2019), Text-classification (Bel et al., 2003; Shi et al., 2010; Mihalcea et al., 2007; Prettenhofer and Stein, 2010; Xu et al., 2016; Chen et al., 2018) etc. However, as far as we are aware, this is the first paper which evaluates the performance of CLT on *Constituency Parsing* task.

In order to generalize a mono-lingual constituency parsing model to multi-lingual settings, we utilize the knowledge of *Language typology* which is available as various typological feature-values in ***World Atlas of Language System (WALS)*** (Haspelmath, 2009) database.

It is observed that CLT based approaches do not perform well if the source and target languages are typologically very distinct (Ruder et al., 2019a). But since *UniRNNG* explicitly models over the typological features (as inputs) and is trained on a sufficiently diverse polyglot corpus, it is comparatively more robust to the typological differences between source and target languages. In other words, once being trained on sufficiently large and typologically diverse corpus it can be applied to any natural-language thus making it *Language-Agnostic*.

Section 2 provides a brief description of *Recurrant Neural Network Grammar (RNNG)* proposed by Dyer' et. al as background work. In section 3 we outline the architecture and intuition behind our proposed **UniRNNG**. Sections 4 and 5 describe the experiments performed and results obtained during the evaluation of proposed model.

2 Background

2.1 Cross-lingual Parsing

Cross-lingual *Model-transfer* approaches to Dependency Parsing such as (Daniel et al., 2017a; Zeman et al., 2018a; Duong et al., 2015; Guo et al., 2016; Vilares et al., 2015; Falenska and Çetinoğlu, 2017; Mulcaire et al., 2019; Vania et al., 2019; Shareghi et al., 2019) involve training a model on high-resource languages and subsequently adapting it to low-resource languages. Participants of CoNLL 2017 shared-task (Daniel et al., 2017b) and CoNLL 2018 shared task (Zeman et al., 2018b) also provide numerous approaches to dependency parsing of low-resource languages. Some approaches such as (Naseem et al., 2012; Täckström et al., 2013; Barzilay and Zhang, 2015; Wang and Eisner, 2016a; Rasooli and Collins, 2017; Ammar, 2016; Wang and Eisner, 2016b) used typological information to facilitate cross-lingual transfer. However all these approach utilise cross-lingual transfer learning for depndency-parsing task while our approach is for the cross-lingual Constituency-parsing/Phrase-parsing.

2.2 Recurrent Neural Network Grammar

RNNGs is a transition based approach to constituency parsing. Transition based parsing approaches reformulate the parsing problem as the task of prediction of best possible action-sequence.

A typical transition-based parser (Jurafsky and Martin, 2019) consists of a *Stack* S which stores the incomplete parse-tree, *Buffer* B which stores the sentence tokens and the set of all possible actions A. At every time-step t, the algorithm chooses the best action $a_t \in A$, given the current state of stack S_t, buffer B_t and history of actions $a_{<t}$. Depending upon the chosen action a_t, the Stack and Buffer are updated accordingly. The process is continued until the Buffer becomes empty and Stack consists of completed parse-tree.

(Dyer et al., 2016) proposed two variants of RNNGs namely *Discriminative* and *Generative* model. The *Discriminative* model computes most probable parse-tree y given the corresponding sentence x whereas the *Generative* RNNG is a language-model that generates sen-

Action	Description
NT(X)	Opens a non-terminal node 'X' and puts it on top of *Stack*. eg: NT(VP)==>(VP
SHIFT	Removes topmost token from the *Buffer* B and pushes onto Stack
REDUCE	Repeatedly pops completed sub-trees or terminal symbols from the stack until an open non-terminal is encountered, and then this open NT is popped and used as the label of a new constituent that has the popped sub-trees as its children. This new completed constituent is pushed onto the stack as a single composite item.

Table 1: Action Set for *Discriminative RNNG* (Dyer et al., 2016)

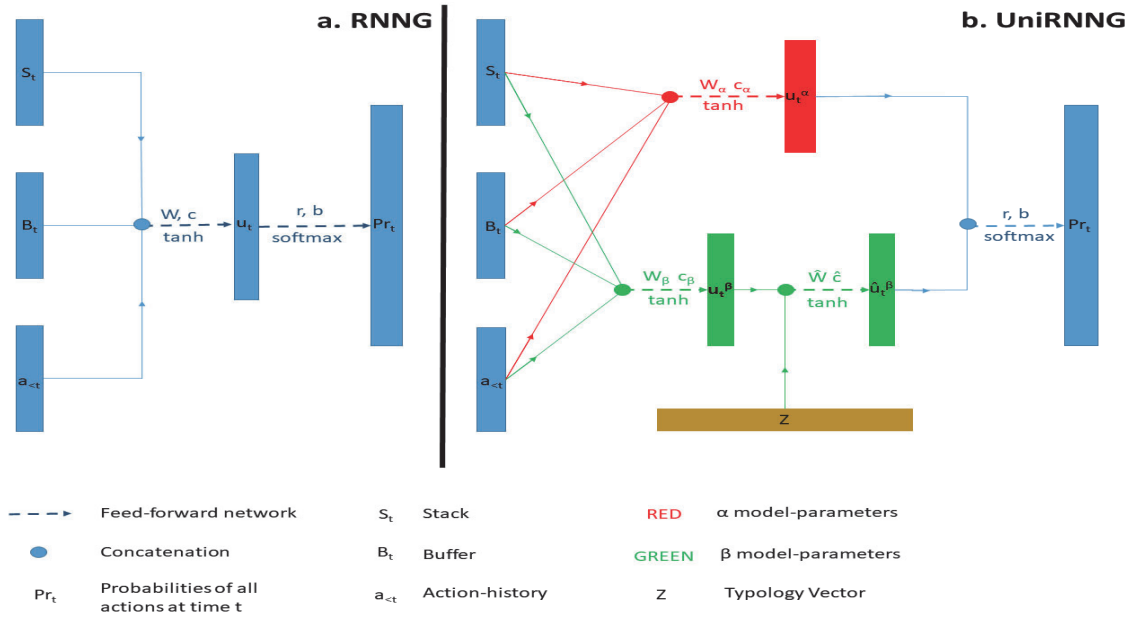

Figure 1: a. Recurrent Neural Network Grammar (RNNG) architecture. b.Universal Recurrent Neural Network Grammar (UniRNNG) architecture.

tence x and y simultaneously. Our proposed **UniRNNG** is a multi-lingual variant of the *Discrimantive* RNNG.

2.2.1 Discrimative RNNG

Table 1 describes the actions within action-set A for the *Discriminative RNNG (DiscRNNG)*. At any time-step t, RNNGs use a stack-LSTM (Dyer et al., 2015) to encode the current state of Stack S_t and use simple RNN to encode the current state of Buffer B_t and action-history $a_{<t}$. Given S_t, B_t and $a_{<t}$, the probability vector P_t comprising probabilities of all actions within A being the appropriate action to be taken at time-step t is computed by applying equation 1.

$$P_t = softmax(r^T u_t + b) \tag{1}$$

Vector u_t is vector representing the entire model-state at time t. u_t is computed by applying equation 2.

$$u_t = tanh(W[S_t; B_t; a_{<t}] + c) \tag{2}$$

Figure 1a depicts the neural-architecture for the entire action-prediction process at any time-step t by the RNNGs.
Given a sentence (token-sequence) x^i and its respective parse-tree y^i as a training example, the action-sequence that generated y^i from x^i can be extracted by depth-first, left-to-right traversal of y^i. The model-parameters are learnt by maximizing the likelihood of this extracted action-sequence for each training example.

3 UniRNNG Model

This section describes our proposed **Universal Recurrent Neural Network Grammar (UniRNNG)**. As being a multi-lingual variant of *DiscRNNG* (section 2.2.1), the

UniRNNG is also a transition based parser consisting of a *Stack* S, *Buffer* B and *action-set* A. At any time-step, the *Stack* stores incomplete parse-tree and *Buffer* stores token-sequence. At each time-step t, model predicts best action $a_t \in A$ given current state of Stack (S_t), Buffer (B_t) and Action-history ($a_{<t}$). Subsequently *Stack* and *Buffer* are updated as S_{t+1} and B_{t+1}, according to action a_t.

3.1 Architecture

Figure 1b depicts the architecture of the **UniRNNG**. At each time-step t the proposed model computes the Stack-encoding S_t, Buffer encoding B_t and action-sequence encoding $a_{<t}$ using stack-LSTM and RNN respectively, in similar way as *DiscRNNG*. (Section 2.2.1). However for **UniRNNG** *Cross-lingual Word-Embeddings* are used instead of Word-Identifier vectors during encoding of Stack and Buffer.

Once having computed S_t, B_t and $a_{<t}$ the model computes two distinct vector-representations of the entire model-state at time t namely α-*vector* (${u^\alpha}_t$) and β-*vector* (${u^\beta}_t$), unlike *DiscRNNG* which computes single representation u_t (equation 2). The ${u^\alpha}_t$ and ${u^\beta}_t$ are computed through equations 3 and 4.

$$ {u^\alpha}_t = tanh(W^\alpha[S_t; B_t; a_{<t}] + c^\alpha) \quad (3) $$

$$ {u^\beta}_t = tanh(W^\beta[S_t; B_t; a_{<t}] + c^\beta) \quad (4) $$

A *typology aware version of* β-*vector* $\hat{u}_t^\beta$ is computed by applying equation 5 (computation simply involves concatenation and dimension reduction through feed-forward network).

$$ \hat{u}_t^\beta = tanh(\hat{W}[{u^\beta}_t; Z] + \hat{c}) \quad (5) $$

Here $Z \in R^{|Z|}$ is a *Linguistic-typology* vector. Each value within Z represents a single typology-feature from *WALS* (Haspelmath, 2009) database having specific value as integer for the language being parsed. Both ${u^\beta}_t$ and $\hat{u}_t^\beta$ have same dimensions i.e. R^d. Final state-representation at time t is given as concatenation of α-*vector* (${u^\alpha}_t$) and *typology aware version of* β-*vector* ($\hat{u}_t^\beta$) as equation 6. Missing features for any language is assigned *zero* indicating no dominant value for it.

$$ u_t = [{u^\alpha}_t; \hat{u}_t^\beta] \quad (6) $$

To summarize *UniRNNG* is very similar to Dyer's *DiscRNNG* 2.2.1 with following modifications.

1. Cross-lingual Word-embeddings are used instead of unique word-identifiers
2. At each time-step t, two distinct model-state representations are computed namely α-*vector* ${u^\alpha}_t$ and β-*vector* ${u^\beta}_t$.
3. Final model-state representation u_t is computed as concatenation of α-*vector* and *typology aware version of* β-*vector*. This is unlike original *DiscRNNG* where u_t is computed directly from S_t, B_t and $a_{<t}$
4. Model is trained on a typologically diverse polyglot corpus.

The proposed architecture is inspired by the *Principle and Parameter framework* (Chomsky, 1993) framework proposed by linguists *Noam Chomsky* and *Howard Lasnik.* (Chomsky, 1993). The central idea behind the PP framework is that a person's syntactic knowledge can be modelled with two formal attributes namely a finite set of fundamental ***Principles*** that are shared by all languages (e.g.: A sentence must always have a subject) and a finite set of ***Parameters*** whose values characterize syntactic variability amongst various languages (eg: *Subject-Verb-Object* (S-V-O) order within a sentence).

Inspired by this PP theory, our proposed *UniRNNG* architecture comprises of distinct α (W^α,c^α) and β (W^β,c^β) parameters to encode the universal and language specific features.

4 Experiments

This section describes the experiments conducted to evaluate the performance of proposed **UniRNNG**. Each experiment comprises of a set of source languages L_s and a single target language l_t.

4.1 Experimental Settings

We evaluated the performance of **UniRNNG** under two experimental setups namely *Few-shot learning* and *Zero-shot learning* setups.

Few-shot Learning (Wang et al., 2019) is applied when only few training examples are

Language	Tree-bank	Family
English	Penn tree-bank (Marcus et al., 1993)	Germanic
Swedish (sd)	Talbanken05 (Nivre et al., 2006)	Germanic
French (fr)	FrenchTreebank (Abeillé et al., 2003)	Romance
Spanish (es)	Spanish UAM Treebank (Moreno et al., 1999)	Romance
Japanese (jp)	Tüba-J/S (Kawata and Bartels, 2000)	Altic
Arabic (ab)	Arabic PENN Treebank (Bies and Maamouri, 2003)	Afro-asiatic
Hungarian (hg)	Hungarian Szeged Treebank (Treebank)	Uralic

Table 2: List of source languages and their corpra used during experimentation. corpra are used to train both *Word-Embeddings* and *Parsers*

Language	Tree-bank	Family
German (de)	Negra Treebank (Skut et al., 1997)	Germanic
Danish (da)	Arboretum Treebank (Bick, 2003)	Germanic
Italian (it)	ISST Treebank (Montemagni et al., 2003)	Romance
Catalan (ct)	Catalan AnCora Treebank (Taulé et al., 2008)	Romance
Korean (kr)	Korean Penn Treebank (Han et al., 2002)	Altic
Heberew (hb)	(Sima' an et al., 2001)	Afro-asiatic
Estonian (est)	Estonian Arborest Treebank (Bick et al.)	Uralic
Hindi (hi)*	Hindi-Urdu Treebank (Bhat et al., 2017)	Indo-aryan
Vietnamese (vt)*	Vietnamese Treebank (Nguyen et al., 2009)	Austroasiatic

Table 3: List of target languages and their corpra used during experimentation. corpra are used to train both *Word-Embeddings* and *Parsers*. * these languages are used only in zero-shot settings

available in the *target language*. In this setup, the cross-lingual models (baseline and **UniRNNG**) are trained on a mixed corpus comprising of source-language sentences (covering over 80% corpus) and few available target language sentences. Hence for *Few-shot Learning* setup $l_t \in L_s$.
Zero-shot Learning (Socher et al., 2013) is applied when no labelled dataset is available in the *target language*. Hence $l_t \notin L_s$.

4.2 Baselines

This section describes the baselines used to compare the performance of our proposed *UniRNNG*.

4.2.1 Mono-lingual Models trained on Sparse Dataset

We used this baseline to compare the performance of our proposed *UniRNNG* only in the *Few-shot* learning settings. As our *UniRNNG* model is intended to be applied for low-resource languages, we compare the performance of it with that of the state-of-the-art mono-lingual models trained on sparse dataset. We experiment with three mono-lingual constituency parsers namely *DiscRNNG* 2.2.1, (Kuncoro et al., 2016) and Transformer (Vaswani et al., 2017).
These models provide over 95% F-Score when trained with sufficiently large dataset. But they would not show such high performance when trained on sparse dataset.

4.2.2 Unsupervised Recurrant Neural Network Grammar (URNNG)

Its a state of the art approach to *unsupervised constituency parsing*. We used this baseline to compare the performance of our proposed *UniRNNG* only in the *Zero-shot* learning settings.

4.2.3 Cross-lingual RNNG Parser trained on single source language (CL-RNNG-Mono)

Its the Dyer's RNNG model (Dyer et al., 2016) with only two modifications. Firstly the *Cross-lingual Word Embeddings* (Ruder et al., 2019b) are used rather than unique word-identifier vectors as used by Dyer et. al. Secondly the model is trained on a single source language *English* (UniRNNGs are trained on poly-

Hyper-parameter	Value
WE dims	768
S_t,B_t,$a_{<t}$ dims	450
${u^\beta}_t$, ${u^\alpha}_t$ dims	450
Dropout prob.	0.01
Bach-size	32
Number of steps per epoch	Size of training corpus / 32
Epochs	150
BERT Model	bert_multi_cased_L-12_H-768_A-12

Table 4: Hyper-parameters

glot corpus) and tested on multiple target language. Within *Few-shot learning*, the training corpus also include small number of labelled target language sentences.

4.2.4 Cross-lingual RNNG Parser trained of multiple source languages (CL-RNNG-Poly)

It is the same model as described in 4.2.3, but trained on a mixed polyglot corpus of high-resource source languages.(*CL-RNNG-Mono* is trained on a single source language *English*). Similar to 4.2.3, a small number of labelled target-language l_t sentences are included as part of the training corpus within the *Few-shot* settings.

4.3 Dataset

Tables 2 and 3 list all the *Source* and *Target* languages as well as their tree-bank corpra used during experimentation. We evaluated our proposed *UniRNNG* model and all the baseline models on each of the target languages listed in Table 3 independently.

As already explained in section 4.1, the *CL-RNNG-Mono* parsers (4.2.3) are always trained on the single source-language *English*, whereas the *CL-RNNG-Poly* and the *UniRNNG* Parsers are always trained on a mixed polyglot corpus (in both *few-shot* and *zero-shot* setups). For each experiment, the source-language training corpus size is always fixed to 700,000 tokens to ensure controlled experiment-settings.

We created the source-language training-corpus for *CL-RNNG-Mono* parsers by randomly sampling sentences from the English-PTB corpus (one at a time), until the token-size becomes approximately equal to 700,000. On the other hand, to create the source-language training-corpus for *CL-RNNG-Poly* and *UniRNNG* models, we randomly sampled sentences from each of the seven source-language corpra listed in table 2 until the token-size becomes approximately equal 100,000, concatenated all these sampled datasets and randomly shuffled the order. Hence all the seven source-languages listed in table 2 are equally represented in the training-corpus for *CL-RNNG-Poly* and *UniRNNG* models.

4.3.1 Short tree-bank corpra

As explained in section 4.1, within *Few-shot learning* settings, only sparse target-language dataset should be used to train both *UniRNNG* and *Baselines*. Hence we extracted a small subset of entire large treebank corpus for each target language listed in table 3.

We extracted this subset by randomly sampling sentences from the target-language tree-bank corpus until the token-size becomes approximately equal to 3000. This is inspired by (Ammar et al., 2016) who used same yard-stick to evaluate their *Multi-lingual Dependency Parser (MALOPA)*. This small target-language language corpus is added to the source-language training corpus for each experiment, within *Few-shot Learning* setup.

4.4 Universal Annotation

There are numerous tree-bank corpra for a diverse range of languages being developed during the years (some listed in Tables 2 and 3). But unlike *Dependency Parsing* tree-banks which are mostly annotated with the *UD Annotations* (McDonald et al., 2013) (for most languages), in case of *Constituency Parsing* various existing tree-bank corpra have their own independent tag annotations, thus making the application of multi-lingual approaches to it as impossible.

However, (Han et al., 2014) proposed a *Universal Phrase tag-set* with 9 common Phrase-tags. Furthermore, (Han et al., 2014) also provides a mapping table to map tags of popular constituency tree-banks (including all treebanks used by us in our experiments) to these *Universal Phrase Tags*.

We used this mapping table to replace all tags within all tree-banks listed in Tables 2 and 3, with the universal tags. Subsequently we trained and evaluated all approaches (including baseline mono-lingual approaches) on these *Universally Tagged* tree-bank versions.

4.5 Cross-Lingual Word Embedding

As our model is a polyglot, we use *Cross-lingual Word-embeddings* during the encoding of Stack and Buffer state at any time-step t. We use a simple *Linear transformation based approach* (Ruder et al., 2019b) to compute such *Cross-lingual Word-embeddings*.

Given two languages l_1 and l_2, the simple *Linear Transformation* based approach first trains the mono-lingual WE for both l_1 and l_2 independently. Subsequently it uses a bi-lingual lexicon to learn a transformation matrix W^{l_1,l_2} to project embeddings of words of l_1 to the embedding-space of l_2 (considering l_2 as reference language).

To ensure that all WE are within same space, we use *English* as reference language. Mono-lingual WE of any other language l are thus transformed into the English space by learning the transformation matrix $W^{l,e}$ from word-pairs extracted from *English-l* bi-lingual lexicon.

We experiment with five common Word-embeddings namely *Skip-gram Word2vec* (Mikolov et al., 2013), *Fast-text* (Grave et al., 2018), Glove (Pennington et al., 2014), ELMo (Peters et al., 2018) and BERT (section 4.5.1). We use bi-lingual seed dictionaries provided by WOLD (Haspelmath and Uri Tadmor, 2009), ASJP (Wichmann and Brown, 2016) and IDS (Key and Comrie, 2015) which are elaborate multi-lingual lexical semantic databases.

4.5.1 BERT Word Embeddings

We computed language-independent BERT-Embeddings to be fed into UniRNNG using pre-trained Multilingual BERT (mBERT) (Wu and Dredze, 2019) model. mBERT is a multilingual variant of original BERT model (Devlin et al., 2018) trained on text from Wikipedia in 104 languages.

The Embeddings are calculated in same way as in (Kondratyuk and Straka, 2019). Given a sentence S, we tokenised the whole sentence using WordPiece tokeniser (Wu et al., 2016). Subsequently we fed this token-sequence into pre-trained mBERT provided by (Turc et al., 2019). Embedding of any word $w \in S$ i.e. e_w is computed by taking average of mBERT outputs of all Wordpiece tokens corresponding to word w.

Thus, mBERT based Word-embeddings do not require any Linear transformation.

4.6 Typology and Hyper-parameters

Table 4 outlines hyper-permeters used during experiments. These values are obtained by minimizing the training loss on *Development* dataset (Dev set) for *Penn Treebank Corpus* (Marcus et al., 1993).

Typology vector Z includes feature-values of all word-order and constituency features in WALS (Haspelmath, 2009) database excluding trivially redundant features as excluded by (Takamura et al., 2016).

5 Results and Inference

Tables 5 outlines results obtained from experiments conducted within the *Few-shot Learning* settings. Best results for *CL-RNNG-Mono*, *CL-RNNG-Poly* and proposed *UniRNNG* models are obtained with BERT Embedding. Table 6 outlines results obtained for experiments conducted under *Zero-shot* learning settings. As we obtained best results with BERT Embeddings within few-shot settings, we experimented with only BERT-embeddings 4.5.1 in *Zero-shot* settings indeed.

As *CL-RNNG-Mono* is trained on the single source language English, it is expected to perform comparatively better on the target languages which are typologically closer to English and poorer on the target languages which are typologically apart from English. On the other hand, *CL-RNNG-Poly* and *UniRNNG* are expected to perform almost uniformly on all the target languages as these are trained on typologically diverse polyglot corpra. These expected trends are in-fact observed in both *Few-shot* and *Zero-shot* learning settings as evident in Tables 5 and 6. Hence for languages Danish (da) and German (de), *Cl-RNNG-Mono* outperformed both *CL-RNNG-Poly* and *UniRNNG* as these languages belong to the same language-family as English namely *Germanic* and are indeed

Model	de	da	it	ct	kr	hb	est
Transformers (Vaswani et al., 2017)	34.34	33.08	34.71	33.74	35.58	35.60	35.57
DiscRNNG 2.2.1	34.49	33.52	35.01	34.15	36.02	35.74	35.94
(Kuncoro et al., 2016)	34.98	33.68	35.53	34.46	36.3	36.42	36.23
CL-RNNG-Mono+Skip-Gram	65.63	70.85	54.59	58.05	22.95	30.44	53.43
CL-RNNG-Mono+Fast-text	67.13	72.55	56.39	60.35	24.75	31.94	55.83
CL-RNNG-Mono+Glove	68.73	74.15	57.29	61.15	25.45	33.84	55.93
CL-RNNG-Mono+ELMo	69.13	74.75	58.49	61.64	26.65	33.94	56.73
CL-RNNG-Mono+BERT	71.03	77.35	60.39	63.05	27.75	39.84	59.93
CL-RNNG-Poly+SkipGram	61.94	62.89	64.0	64.53	61.88	63.19	62.76
CL-RNNG-Poly+Fast-text	63.57	64.51	65.78	66.53	64.3	64.84	65.55
CL-RNNG-Poly+Glove	65.1	66.17	66.5	67.4	64.72	66.59	65.51
CL-RNNG-Poly+ELMo	65.48	66.86	67.61	68.16	65.89	66.64	66.01
CL-RNNG-Poly+BERT	67.48	69.41	69.55	70.46	69.18	69.88	69.19
UniRNNG+SkipGram	64.92	65.95	66.79	67.35	65.05	66.24	65.83
UniRNNG+Fast-text	66.42	67.65	68.59	69.64	67.05	67.74	68.23
UniRNNG+Glove	68.03	69.25	69.49	70.45	67.55	69.64	68.33
UniRNNG+ELMo	68.42	69.85	70.69	70.94	68.75	69.74	69.13
UniRNNG+BERT	70.33	72.44	72.59	73.35	71.85	72.64	72.33

Table 5: F1 Score in *Few-shot* learning settings. *Top:* Results for supervised approaches trained on sparse dataset. *Middle:* Results for baseline Cross-lingual Transfer Parser (CLT-P). *Bottom:* Results for proposed **UniRNNG**

Model	de	da	it	ct	kr	hb	est	hi	vt
URNNG (Kim et al., 2019)	11.84	11.58	10.53	12.43	9.97	10.46	8.52	9.36	3.12
CL-RNNG-Mono+BERT	68.13	70.94	61.99	56.85	20.91	27.82	52.61	48.66	37.61
CL-RNNG-Poly+BERT	64.43	64.13	64.5	66.37	63.32	64.99	63.5	56.2	57.21
UniRNNG+BERT	67.62	67.03	67.19	69.14	66.25	68.14	66.63	59.23	60.11

Table 6: F1 Score in *Few-shot* learning settings.

typologically very close to English. Whereas, on the other five target languages which are typologically and genealogically distinct from the source language English namely Italian (it), Catalan (ct), Estonian (est), Heberew (hb) and Korean (kr), it under-performed *CL-RNNG-Poly*.

Based on these observed trends we can infer that the polyglot training training increases the Cross-lingual transferring ability of the RNNG based Constituency Parser to a typologically distinct and unseen target language as it allows the model to better generalize over a diverse set of languages.

In both *Few-shot* and *Zero-shot* settings, *UniRNNGs* significant outperformed *CL-RNNG-Poly* on all the seven target languages namely Danish (da), German (de), Italian (it), Catalan (ct), Estonian (est), Heberew (hb) and Korean (kr) as evident in Tables 5 and 6. Hence it can be inferred inducing linguistic typology indeed leads to further improvement in Cross-lingual transferring ability of the RNNG based Constituency Parser to a typologically distinct and unseen target language.

Furthermore, in *zero-shot* learning settings, we evaluated our models on two additional target languages namely *Hindi* and *Vietnamese* (rightmost column in table 6). Languages *Hindi* and *Vietnamese* belong to linguistic families *Indo-aryan* and *Austro-asiatic* respectively. None of the source languages listed in Table 2 belong to these linguistic families. Thus languages *Hindi* and *Vietnamese* are typologically very distant form all the source languages in the polyglot training corpus of *UniRNNGs*. Hence scores obtained on these languages indicate true Language Agnostic nature of **UniRNNG** architecture.

Although the performance of **UniRNNG** for these two languages is comparatively lower than its performance on other target languages listed in table 3, yet this improved performance as compared to *CL-RNNG-Mono* and *CL-RNNG-Poly* provide even stronger evidence that **UniRNNG** architecture is more robust to typologically distinct unseen target languages than *CL-RNNG-Poly*. In other words, once trained on significantly diverse polyglot corpus, **UniRNNG** is *Language-Agnostic*.

6 Conclusion

In this work, we proposed and evaluated *Universal Recurrent Neural Network Grammar (UniRNNG)* which is a multilingual variant of Dyer's RNNG model. The architecture of *UniRNNG* is inspired by *Principles and Parameters* theory proposed by linguist Noam Chomsky. We evaluated the performance of *UniRNNG* in both *Few-shot* and *Zero-shot* learning settings. Results show that the *UniRNNGs* outperformed all baseline approaches for most of the target languages for which these are tested. As far as we are aware, this is the first paper which evaluated the performance of Cross-lingual Transfer Parsing for *Constituency Parsing* task.
Future work, would involve exploring the changes in performances of baseline and *UniRNNG* models with the varying degree of diversity in the training corpus.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for french. In *Treebanks*, pages 165–187. Springer.

Waleed Ammar. 2016. *Towards a Universal Analyzer of Natural Languages*. Ph.D. thesis, Ph. D. thesis, Google Research.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Mark C Baker. 2008. *The atoms of language: The mind's hidden rules of grammar*. Basic books.

Regina Barzilay and Yuan Zhang. 2015. Hierarchical low-rank tensors for multilingual transfer parsing. Association for Computational Linguistics.

Nuria Bel, Cornelis HA Koster, and Marta Villegas. 2003. Cross-lingual text categorization. In *International Conference on Theory and Practice of Digital Libraries*, pages 126–139. Springer.

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, et al. 2017. The hindi/urdu treebank project. In *Handbook of Linguistic Annotation*, pages 659–697. Springer.

Eckhard Bick. 2003. Arboretum, a hybrid treebank for danish. In *Proceedings of TLT 2003 (2nd Workshop on Treebanks and Linguistic Theory, Växjö*, pages 9–20.

Eckhard Bick, Heli Uibo, and Kadri Muischnek. Preliminary experiments for a cg-based syntactic tree corpus of estonian.

Ann Bies and Mohamed Maamouri. 2003. Penn arabic treebank guidelines. *Draft: January*, 28:2003.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

Noam Chomsky. 1986. *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Noam Chomsky. 1993. *Lectures on government and binding: The Pisa lectures*. 9. Walter de Gruyter.

Guglielmo Cinque and Luigi Rizzi. 2010. The cartography of syntactic structures. *Oxford Handbook of linguistic analysis*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Vivian Cook and Mark Newson. 2014. *Chomsky's universal grammar*. John Wiley & Sons.

Zeman Daniel, Popel Martin, Straka Milan, Hajic Jan, Nivre Joakim, Ginter Filip, Luotolahti Juhani, Pyysalo Sampo, Petrov Slav, Potthast Martin, et al. 2017a. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, volume 1, pages 1–19. Association for Computational Linguistics.

Zeman Daniel, Popel Martin, Straka Milan, Hajic Jan, Nivre Joakim, Ginter Filip, Luotolahti Juhani, Pyysalo Sampo, Petrov Slav, Potthast Martin, et al. 2017b. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, volume 1, pages 1–19. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. 2019. Meemi: A simple method for post-processing cross-lingual word embeddings. *arXiv preprint arXiv:1910.07221*.

Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. 2015. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 845–850.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. *arXiv preprint arXiv:1505.08075*.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. 2016. Recurrent neural network grammars. *arXiv preprint arXiv:1602.07776*.

Agnieszka Falenska and Özlem Çetinoğlu. 2017. Lexicalized vs. delexicalized parsing in low-resource scenarios. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 18–24.

Janet Dean Fodor. 2001. Setting syntactic parameters.

Janet Dean Fodor and William Gregory Sakas. 2004. Evaluating models of parameter setting. In *Proceedings of the 28th annual boston university conference on language development*, volume 1, pages 1–27. Cascadilla Press Somerville, MA.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. 2016. A representation learning framework for multi-source transfer parsing. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Aaron Li-Feng Han, Derek F Wong, Lidia S Chao, Yi Lu, Liangye He, and Liang Tian. 2014. A universal phrase tagset for multilingual treebanks. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 247–258. Springer.

Chunghye Han, Narae Han, Eonsuk Ko, and Martha Palmer. 2002. Korean treebank: Development and evaluation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*.

Martin Haspelmath. 2009. *The typological database of the World Atlas of Language Structures*. Berlin: Walter de Gruyter.

Martin Haspelmath and editors Uri Tadmor. 2009. WOLD.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964*.

Daniel Jurafsky and James H. Martin. 2019. Transition-based dependency parsing (section 15.4). In *Speech and Language Processing (3rd Edition draft)*, chapter 15, pages 6–17.

Yasuhiro Kawata and Julia Bartels. 2000. Stylebook for the japanese treebank in verbmobil. In *Verbmobil-Report 240, Seminar für Sprachwissenschaft, Universität Tübingen*.

Richard Kayne. 2012. Some notes on comparative syntax, with special reference to english and french. In *The Oxford handbook of comparative syntax*. Oxford University Press.

Mary Ritchie Key and Bernard Comrie. 2015. IDS.

Yoon Kim, Alexander M Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. 2019. Unsupervised recurrent neural network grammars. *arXiv preprint arXiv:1904.03746*.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. *arXiv preprint arXiv:1904.02099*.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A Smith. 2016. What do recurrent neural network grammars learn about syntax? *arXiv preprint arXiv:1611.05774*.

Chia-Hsuan Lee and Hung-Yi Lee. 2019. Cross-lingual transfer learning for question answering. *arXiv preprint arXiv:1907.06042*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. Xqa: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 976–983.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, et al. 2003. The italian syntactic-semantic treebank: Architecture, annotation, tools and evaluation.

Antonio Moreno, Susana López, and Manuel Alcántara. 1999. Spanish tree bank: Specifications, version 5. *Technical paper*.

Phoebe Mulcaire, Jungo Kasai, and Noah A Smith. 2019. Low-resource parsing with crosslingual contextualized representations. *arXiv preprint arXiv:1909.08744*.

Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. The Association for Computational Linguistics.

Phuong Thai Nguyen, Xuan Luong Vu, Thi Minh Huyen Nguyen, Hong Phuong Le, et al. 2009. Building a large syntactically-annotated corpus of vietnamese.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *LREC*, pages 1392–1395.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Steven Pinker. 1995. *The language instinct: The new science of language and mind*, volume 7529. Penguin UK.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 1118–1127.

Mohammad Sadegh Rasooli and Michael Collins. 2017. Cross-lingual syntactic transfer with limited resources. *Transactions of the Association for Computational Linguistics*, 5:279–293.

Ian Roberts and Anders Holmberg. 2005. On the role of parameters in universal grammar: A reply to newmeyer. *Organizing Grammar: Linguistic Studies in Honor of Henk van Riemsdijk. Berlin: Mouton de Gruyter*, pages 538–553.

Sebastian Ruder, Anders Søgaard, and Ivan Vulić. 2019a. Unsupervised cross-lingual representation learning. In *Proceedings of ACL 2019, Tutorial Abstracts*, pages 31–38.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019b. A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Ehsan Shareghi, Yingzhen Li, Yi Zhu, Roi Reichart, and Anna Korhonen. 2019. Bayesian learning for neural dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3509–3519.

Lei Shi, Rada Mihalcea, and Mingjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1057–1067.

Ur Shlonsky. 2010. The cartographic enterprise in syntax. *Language and linguistics compass*, 4(6):417–429.

Khalil Sima' an, Alon Itai, Yoad Winter, Alon Altman, and Noa Nativ. 2001. Building a tree-bank of modern hebrew text. *Traitement Automatique des Langues*, 42(2):247–380.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering. *arXiv preprint arXiv:1905.11471*.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. *arXiv preprint cmp-lg/9702004*.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.

Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2016. Discriminative analysis of linguistic features for typological study. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 69–76.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*.

Michael Tomasello. 2005. Beyond formalities: The case of language acquisition. *The Linguistic Review*, 22(2-4):183–197.

Hungarian Szeged Treebank. Szeged treebank 2.0: A hungarian natural language database with detailed syntactic analysis. *Hungarian linguistics at the University of Szeged*.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Clara Vania, Yova Kementchedjhieva, Anders Søgaard, and Adam Lopez. 2019. A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. *arXiv preprint arXiv:1909.02857*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

David Vilares, Carlos Gómez-Rodríguez, and Miguel A Alonso. 2015. One model, two languages: training bilingual parsers with harmonized treebanks. *arXiv preprint arXiv:1507.08449*.

Dingquan Wang and Jason Eisner. 2016a. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.

Dingquan Wang and Jason Eisner. 2016b. The galactic dependencies treebanks: Getting more data by synthesizing new languages. *Transactions of the Association for Computational Linguistics*, 4:491–505.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2019. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*.

Eric W. Holman Wichmann, Søren and Cecil H. Brown. 2016. The ASJP Database (version17).

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Ruochen Xu, Yiming Yang, Hanxiao Liu, and Andrew Hsi. 2016. Cross-lingual text classification via model translation with limited dictionaries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 95–104.

Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018a. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.

Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018b. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.