SVEUČILIŠTE U SPLITU
FAKULTET ELEKTROTEHNIKE, STROJARSTVA I BRODOGRADNJE

POSLIJEDIPLOMSKI DOKTORSKI STUDIJ
ELEKTROTEHNIKE I INFORMACIJSKE TEHNOLOGIJE

KVALIFIKACIJSKI ISPIT

COMPOSITIONAL GENERALIZATION IN SPEECH RECOGNITION FOR MORPHOLOGICALLY RICH LANGUAGES

Lucija Bročić

Split, rujan 2026.

UNIVERSITY OF SPLIT
FACULTY OF ELECTRICAL ENGINEERING, MECHANICAL ENGINEERING AND NAVAL ARCHITECTURE

POSTGRADUATE STUDY OF ELECTRICAL ENGINEERING AND INFORMATION TECHNOLOGY

QUALIFYING EXAM

COMPOSITIONAL GENERALIZATION IN SPEECH RECOGNITION FOR MORPHOLOGICALLY RICH LANGUAGES

Lucija Bročić

Split, September 2026

TABLE OF CONTENTS

1. INTRODUCTION 1
2. BACKGROUND 3
2.1. Compositionality in language and cognition 3
2.2. Automatic speech recognition: core concepts 4
2.3. Morphological richness and the scope of South Slavic 6
2.4. The intersection: Compositional generalization in speech 7
3. METHODOLOGY 8
4. COMPOSITIONAL GENERALIZATION IN NLP 11
4.1. Benchmarks and diagnostic methods 12
4.2. Architectural approaches 15
4.3. Training and data-augmentation approaches 16
4.4. Compositional generalization in semantic parsing, text-to-SQL, and knowledge-based question answering 18
4.5. Compositional generalization in LLMs 19
4.6. Morphology, multilingual settings, and the path toward speech 20
5. CROATIAN AND SOUTH SLAVIC ASR 23
5.1. Speech corpora and language resources 24
5.1.1. Large multilingual initiatives 25
5.1.2. Broadcast news, parliamentary, and large-scale spoken corpora 26
5.1.3. Domain-Specific and Read-Speech Corpora 27
5.1.4. Pronunciation Lexicons and Prosodic Annotation Resources 27
5.1.5. Specialized, Emotional, and Spontaneous Speech Corpora 28
5.2. Statistical ASR systems 29
5.2.1. Early telephone-based recognition and the SpeechDat foundation 29
5.2.2. The morphological challenge: sub-word modeling and lexical adaptation 31
5.2.3. Advancing acoustic modeling 31
5.2.4. Feature optimization and dimensionality reduction 32
5.2.5. Cross-lingual transfer and resource-efficient bootstrapping 33
5.2.6. Robustness and spontaneous speech phenomena 33
5.2.7. Transition to neural architectures 34
5.3. Neural and end-to-end ASR 34

5.3.1. From GMM-HMM Hybrids to deep neural acoustic models 34
5.3.2. Neural language modelling and morphological integration 36
5.3.3. Transfer Learning and domain adaptation 36
5.3.4. End-to-end architectures with CTC-based training 37
5.3.5. Conformer and transformer end-to-end systems 38
5.4. Challenges in South Slavic ASR 39
5.4.1. Linguistic complexity: morphology, prosody, language modelling 39
5.4.2. Resource and acoustic constraints: data scarcity and acoustic variability 41
6. SYNTHESIS OF RESULTS 43
7. CONCLUSION 48
LITERATURE 50
ABBREVIATION LIST 65
ABSTRACT 67
SAŽETAK 68

1. INTRODUCTION

Human language works on a simple but powerful principle called compositionality. We constantly understand and create completely new statements just by recombining familiar words and structures. A child who hears a brand-new sentence doesn't get confused, as they easily figure out what it means because they already understand the individual words and how they fit together. In artificial intelligence (AI), this ability is commonly referred to as *compositional generalization* and is often considered an essential aspect of systematic language understanding.

Despite the remarkable success of neural sequence-to-sequence models, evidence accumulated over the last decade suggests that high predictive accuracy does not necessarily imply systematic generalization. Benchmarks such as SCAN [1] and COGS [2] have demonstrated that models achieving near-perfect performance on standard in-distribution test sets can fail dramatically when confronted with novel but well-formed combinations of familiar elements. These findings have motivated extensive research into the compositional limitations of transformer-based and other neural architectures, suggesting that they frequently rely on distributional patterns rather than learning robust representations of underlying structure.

At the same time, most studies of compositional generalization have been conducted in text-only settings and have focused primarily on English and other morphologically simple languages. Far less attention has been paid to speech technologies and to morphologically rich languages, where extensive inflection and derivation create large vocabularies and substantial surface variation. These characteristics are particularly pronounced in Croatian and related Slavic languages, raising the question of whether modern transformer-based automatic speech recognition (ASR) systems and large pretrained language models can exploit morphological structure to generalize beyond previously observed word forms.

Answering this question would require bringing together several research areas that have so far developed mostly in isolation: compositionality in language and cognition, neural sequence modeling, and speech recognition for low-resource, morphologically rich languages. Because few, if any, existing studies bridge these areas directly, this review first maps each of them separately, examining what is currently known about compositional generalization on one side, and about Croatian and Slavic ASR resources on the other, as a first step toward identifying where and how such a convergence might take place.

To address this gap, this paper presents a systematic review of research on compositional generalization and its relevance to speech recognition in morphologically rich languages. Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [3], the review synthesizes findings from both the broader compositional generalization literature and existing research on Croatian and Slavic ASR. In particular, the review seeks to answer the following research questions (RQs):

RQ1: What approaches and benchmarks have been proposed for studying compositional generalization in text and spoken language, and what are their limitations?

RQ2: What speech resources, ASR models, and evaluation methodologies exist for Croatian and related low-resource Slavic languages?

The remainder of this review is organized as follows: Section 2 provides the background, covering compositionality in language and cognition, core concepts in ASR, the scope and morphological typology of South Slavic languages, and their intersection in compositional generalization in speech; Section 3 describes the review methodology, including the search strategy and screening procedure applied across databases, with the domain-specific eligibility criteria for each literature strand presented alongside the corresponding results in Sections 4 and 5; Section 4 presents the results on compositional generalization in NLP, including benchmarks and diagnostic methods, architectural and training solutions, transformers and LLMs, compositional generalization in structured and spoken language understanding and morphological compositional generalization in multilingual and low-resource contexts; Section 5 reviews Croatian and South Slavic ASR research, covering speech corpora and language resources, statistical ASR systems, neural and end-to-end approaches, and the challenges specific to South Slavic ASR; Section 6 synthesizes the findings across both areas of literature and discusses their implications for the doctoral research; and Section 7 concludes the review.

2. BACKGROUND

2.1. Compositionality in language and cognition

The principle of compositionality is a foundational assumption in formal linguistics and philosophy of language, commonly attributed to Frege and developed in formal semantics tradition, including Montague grammar [4] and related work on direct compositionality [5]. It states that the meaning of a complex expression is derived from the meanings of its parts and the compositional rules that combine them. This assumption provides a basis for explaining the productivity of human language, i.e., how humans can produce and understand an unlimited number of sentences using finite vocabulary and a set of grammatical rules.

Mathematically, within Montague grammar, compositionality can be understood as a homomorphism from the syntactic algebra of expressions to the semantic algebra of meanings [4]. Building on this idea, modern formalizations, such as that of An and Du [6], introduce a finite set of primitives P, and a syntactic composition operator $\circ$ defined by a grammar G. These form an expression space $\mathcal{E}$ generated from these primitives by repeated application of the operator. Under this view, each expression $e \in \mathcal{E}$ is assigned a semantic interpretation $[\![e]\!]$ in a meaning space $\mathcal{S}$, where $\mathcal{S}$ is a semantic algebra with composition operator $*$. The mapping $[\![\cdot]\!]: \mathcal{E} \to \mathcal{S}$ is compositional if for all $e_1, e_2 \in \mathcal{E}$,

$$[\![e_1 \circ e_2]\!] = [\![e_1]\!] * [\![e_2]\!].$$

In other words, the semantic mapping should behave consistently with the grammar, so that the meaning of a complex expression can be derived systematically from the meanings of its constituents rather than learned as an isolated form. This perspective is particularly useful because it connects the classical linguistic notion of compositionality with a precise algebraic criterion that can be measured, analyzed, and potentially enforced in neural representations.

Within cognitive science, compositionality has been used to model how humans generalize knowledge to novel situations, suggesting that structured symbolic representations play a key role in systematic generalization. This view was most influentially articulated by Fodor and Pylyshyn [7], who argued that the compositional structure of language reflects a deeper compositional structure of thought, and that symbolic representations with explicit constituent structure are necessary to explain systematic generalization. According to them, a system capable of understanding “John loves Mary” must, by the same mechanism, be able to understand “Mary loves John”; otherwise, the two cases would amount to unrelated learned behaviors. Their 1988 critique of connectionism framed this as a dilemma: either distributed neural representations cannot account for systematicity and productivity, or, if they can, they must be implementing symbolic structure under the hood. This challenge prompted decades of research and refinement, including neural approaches to compositional structure. The rise of large-scale neural sequence-to-sequence models made this question empirically testable, since these models are precisely the distributed representations Fodor and Pylyshyn's dilemma was about.

In modern machine learning, this debate reappears as *compositional generalization* (CG), referring to a model’s ability to recombine known elements in unseen ways while

preserving meaning. CG is often evaluated through benchmarks that deliberately test this ability, most notably SCAN, which showed that sequence-to-sequence models can fail on systematic recombination, and COGS, which distinguishes lexical from structural generalization. However, recent work has shown that compositional generalization is not always defined consistently across studies, and that benchmark design can strongly affect what is being measured [8], [9]. Because definitions of compositional generalization vary, it is useful to view it as a set of related behaviors rather than a single ability. Hupkes et al. [10] identify five aspects and test each separately: systematicity, recombining known parts and rules in unseen ways; productivity, extending predictions to sequences longer than those seen during training; substitutivity, preserving predictions when a constituent is replaced by a synonymous one; localism, whether composition depends on local constituents or the entire input; and overgeneralization, whether a model applies a general rule or memorizes exceptions when both fit the training data. Evaluating recurrent, convolutional, and transformer sequence-to-sequence models on a synthetic dataset generated by a probabilistic context-free grammar, they find that no architecture performs consistently well across all five aspects, with different architectures showing different weaknesses. Despite this framework, the underlying debate remains unresolved. Surveying researchers active in the area, McCurdy et al. [11] find broad agreement that compositional behavior is not achieved by current models and skepticism that scale alone will achieve it, but no consensus on how it should be evaluated or how the field should proceed.

An important refinement of this debate is that productivity and systematic compositionality are distinct properties rather than equivalent properties. As Baroni [12] notes, modern neural networks can generalize to novel grammatical sentences, suggesting they capture aspects of hierarchical syntax. However, they perform poorly on tasks such as SCAN that require systematically applying learned rules in new contexts. In other words, a model may be productive without being systematically compositional, relying on similarity-based generalization rather than algebraic composition rules.

2.2. Automatic speech recognition: core concepts

Because this review connects compositional generalization to automatic speech recognition (ASR) for morphologically rich languages, this section first introduces ASR as the task of converting spoken language into written text and outlines its core modeling assumptions. ASR systems typically map an acoustic signal, segmented into short frames, to a sequence of linguistic units such as phonemes, subwords, or words, while also relying on a language model to help resolve ambiguity in recognition. Over time, ASR has evolved from classical statistical pipelines based on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) to neural and end-to-end approaches that learn the mapping more directly from data. This section then introduces the modeling paradigms and evaluation metrics referenced throughout the rest of this work.

Hidden Markov Models (HMMs). For several decades, the dominant statistical framework for ASR was the Hidden Markov Model, which represents speech as a sequence of hidden states (typically corresponding to phonetic sub-units) generating observable acoustic

feature vectors, with state transitions governed by a Markov process. This framework and its application to speech recognition were canonically formalized by Rabiner in his widely cited 1989 tutorial [13], which remains the standard reference for the forward-backward and Viterbi algorithms underlying HMM-based recognizers. HMMs allow the temporal variability of speech to be modeled probabilistically, but they require a separate model of the probability that a given acoustic observation was generated by a given state, which is the emission probability.

Gaussian Mixture Models (GMMs). In the classical HMM-GMM paradigm, these emission probabilities were modeled using Gaussian Mixture Models: weighted sums of multivariate Gaussian distributions fit to the acoustic feature space, most commonly Mel-Frequency Cepstral Coefficients (MFCCs), the perceptually motivated acoustic representation introduced by Davis and Mermelstein [14]. Their 1980 study showed MFCCs outperform linear-frequency and linear-prediction cepstral representations for monosyllabic word recognition. The HMM-GMM combination formed the standard ASR architecture from the 1990s through the early 2010s and underlies the majority of the classical South Slavic systems reviewed in Section 5.2, where questions such as covariance structure, feature dimensionality reduction (via LDA/HLDA), and phonetic decision-tree state tying were central engineering concerns.

From hybrid to end-to-end neural ASR. Neural networks were first incorporated by replacing GMM emission estimators with deep neural networks (DNNs) while retaining the HMM backbone — the hybrid DNN-HMM paradigm. This transition was catalyzed by Hinton et al.'s 2012 overview [15], representing the shared findings of four research groups, which demonstrated that DNNs producing posterior probabilities over HMM states substantially outperformed GMMs across multiple speech recognition benchmarks. This hybrid paradigm is discussed for Serbian and Slovenian systems in Section 5.3.1. Subsequently, end-to-end approaches emerged that dispense with an explicit HMM state sequence altogether. Connectionist Temporal Classification (CTC) is the foundational end-to-end training criterion, introduced by Graves et al. in 2006 [16]: it adds a “blank” symbol and marginalizes over all possible alignments between an input acoustic sequence and an output label sequence, allowing a recurrent neural network to be trained directly on unsegmented speech-transcript pairs without frame-level state labels. In their original TIMIT experiments, CTC-trained bidirectional LSTM networks outperformed standard HMMs and HMM-neural network hybrids without requiring presegmented training data. CTC and its successors (attention-based encoder-decoders, and later transformer-based architectures) removed the need for hand-engineered pronunciation lexicons and phonetic decision trees that dominated the classical era. However, as Sections 5.3 and 5.4, lexicon-free modeling creates new challenges for morphologically rich languages with large, productive vocabularies. Because these models must learn how stems and inflectional suffixes combine from the training data rather than from a fixed dictionary, their ability to generalize compositionally becomes crucial for recognizing unseen word forms.

Evaluation metrics: WER and CER. ASR system performance is commonly evaluated using Word Error Rate (WER), which quantifies the minimum number of edit

operations required to transform the recognized word sequence into the reference transcription. Formally, WER is defined as

$$WER = \frac{S + D + I}{N},$$

where S, D, and I denote the minimum number of substitutions, deletions, and insertions, respectively, obtained by aligning a reference sequence and a hypothesis sequence using the Levenshtein edit-distance algorithm, with N denoting the number of words in the reference transcription. Character Error Rate (CER) is defined analogously at the character level, using the same edit operations but with characters as the basic unit.

Consequently, CER provides a more fine-grained measure of transcription quality that is particularly informative for morphologically rich languages, where a single character error can change the grammatical form while preserving most of the word. As a result, WER and CER may differ substantially: an incorrectly recognized suffix counts as a full word error in WER but only a few character errors in CER. This divergence is much more pronounced in South Slavic languages than in English and, as shown in Section 5.3.1, can be roughly fourfold, directly reflecting morphological rather than acoustic errors. Both WER and CER are based on the edit-distance formulation introduced by Levenshtein in 1966 [17], which established the dynamic programming algorithm for measuring insertions, deletions, and substitutions that these metrics still use today.

2.3. Morphological richness and the scope of South Slavic

Morphologically rich languages (MRLs) express grammatical categories such as case, number, gender, tense, and aspect through extensive inflection and derivation rather than through word order or function words alone. A single lemma can have numerous distinct word forms (up to roughly 18 for nouns and 100 for verbs in Slovenian, for example [18]), producing vocabulary growth rates far exceeding those of English and correspondingly severe out-of-vocabulary (OOV) problems for both language modeling and lexicon construction. In ASR specifically, this morphology–vocabulary interaction complicates the mapping from acoustics to words: a system may correctly recognize the acoustic-phonetic content of an utterance yet still produce a WER error because it selected the wrong inflectional suffix, as documented repeatedly for Serbian and Slovenian systems.

This review focuses on Croatian, Serbian, Slovenian, and Bosnian as its primary South Slavic languages of interest. This choice reflects the review's two main objectives. First, these languages are historically closely related and share many morphological and phonological features, including case, gender, and aspect marking, as well as broadly similar phonemic inventories. As a result, resource-development strategies, subword modeling techniques, and cross-lingual acoustic transfer methods developed for one language may also apply to the others, as discussed in Sections 5.2 and 5.3. Second, the systematic search found much less literature on Bulgarian and Macedonian. Although also South Slavic, these languages differ more substantially in their morphology, including the loss of case marking and the presence of definite articles, as well as verbal evidentiality in Bulgarian [19], [20]. They were therefore excluded from the empirical scope of this review.

2.4. The intersection: Compositional generalization in speech

Although compositionality and morphological richness are well-studied in NLP and cognitive science, their interaction in speech recognition has only recently begun to receive attention. Work on spoken language understanding and related tasks shows that state-of-the-art models often rely on spurious correlations and struggle to generalize compositionally to novel combinations of known slots or to longer utterances, despite strong in-domain performance [21]. In parallel, studies of morphological compositional generalization in large language models indicate that models can handle many seen morphological patterns but still exhibit sharp drops in accuracy when asked to apply known morphological processes to novel roots or more complex combinations, suggesting a lack of fully systematic behavior [22].

For speech technologies in morphologically rich, low-resource languages, an important question is whether ASR systems can generalize to unseen but linguistically valid combinations of morphemes and stems, despite limited training data and high word-form variability. Standard ASR evaluations mainly use WERs and CERs, neither of which shows whether a model has learned compositional structure or is simply memorizing frequent word forms.

This literature review addresses this gap by examining research on compositional generalization, morphology-aware modeling, and speech recognition, with a focus on whether compositional structure could improve robustness and data efficiency for morphologically rich languages. Within the scope of the databases and search terms used in this review (Section 3), no study was found that addressed all three aspects of the present work: compositional generalization, speech recognition, and morphologically rich languages. The review therefore brings together findings from these areas and discusses how existing approaches to compositional generalization could motivate future research on speech recognition in morphologically rich languages, including possible evaluation methods and modeling strategies.

3. METHODOLOGY

This review follows the PRISMA 2020 guidelines, which structure a systematic review around its rationale and research questions, database sources and search strategy, eligibility criteria, screening procedure, and synthesis of included studies. As the rationale and research questions were established in the preceding sections, this section focuses on the database sources and search strategies used to identify works within the scope of this review.

Table 3.1. Search strings and number of records retrieved per database for Query 1 (compositional generalization terms combined with NLP/speech domain terms).

Database	Query	#
Web of Science (WosCC)	TS=((“compositional generali?ation” OR “compositionality” OR “systematic generali?ation” OR “systematicity”) AND (“speech recognition” OR “ASR” OR “spoken language understanding” OR “SLU” OR “natural language processing” OR “NLP” OR “natural language understanding” OR “NLU”))	148
Scopus	TITLE-ABS-KEY((“compositional generalization” OR “compositional generalization” OR “compositionality” OR “systematic generalization” OR “systematic generalization” OR “systematicity”) AND (“speech recognition” OR “ASR” OR “spoken language understanding” OR “SLU” OR “natural language processing” OR “NLP” OR “natural language understanding” OR “NLU”))	378
IEEE Xplore	(“Abstract”: ”compositional generalization” OR “Abstract”: ”compositional generalization” OR “Abstract”: “compositionality” OR “Abstract”: “systematic generalization” OR “Abstract”: “systematic generalization” OR “Abstract”: “systematicity”) AND (“Abstract”: “speech recognition” OR “Abstract”: “ASR” OR “Abstract”: “spoken language understanding” OR “Abstract”: “SLU” OR “Abstract”: “natural language processing” OR “Abstract”: “NLP” OR “Abstract”: “natural language understanding” OR “Abstract”: “NLU”)	9

Three scientific databases indexing peer-reviewed journals and conference proceedings were searched: Web of Science (Core Collection), Scopus, and IEEE Xplore (last accessed May 2026). The search strategy was structured around two complementary queries. Query 1

(Q1) combined terms related to compositional generalization, systematicity, and compositionality with NLP and speech technology domain terms. Query 2 (Q2) combined Slavic language identifiers (including Croatian, Serbian, and Slovenian), with terms related to automatic speech recognition, speech corpora, and acoustic modelling. Both queries were applied to titles, abstracts, and author keywords. The exact query strings, and the number of records retrieved per database are presented in Table 3.1 for Q1 and Table 3.2 for Q2.

Table 3.2. Search strings and number of records retrieved per database for Query 2 (South Slavic language identifiers combined with ASR/speech-resource terms).

Database	Query	#
Web of Science (WosCC)	TS=((“Croatian” OR “South Slavic” OR “Slovenian” OR “Serbian” OR “Bosnian” OR “low-resource Slavic”) AND (“speech recognition” OR “ASR” OR “spoken language” OR “speech corpus” OR “acoustic model” OR “speech dataset” OR “spoken digit*” OR “spoken number*”))	136
Scopus	TITLE-ABS-KEY((“Croatian” OR “South Slavic” OR “Slovenian” OR “Serbian” OR “Bosnian” OR “low-resource Slavic”) AND (“speech recognition” OR “ASR” OR “spoken language” OR “speech corpus” OR “acoustic model” OR “speech dataset” OR “spoken digit” OR “spoken number”))	351
IEEE Xplore	(“Abstract”: “Croatian” OR “Abstract”: “South Slavic” OR “Abstract”: “Slovenian” OR “Abstract”: “Serbian” OR “Abstract”: “Bosnian” OR “Abstract”: “low-resource Slavic”) AND (“Abstract”: “speech recognition” OR “Abstract”: “ASR” OR “Abstract”: “spoken language” OR “Abstract”: “speech corpus” OR “Abstract”: “acoustic model” OR “Abstract”: “speech dataset” OR “Abstract”: “spoken digit” OR “Abstract”: “spoken number”)	46

The database search using Q1 yielded in total 535 records. Using Rayyan screening tool [23], 147 duplicates were automatically detected and removed. Of the remaining 388 records, 33 additional records were excluded after manual inspection: 17 records corresponding to entire conference proceeding volumes, 2 books, and 14 book chapters. This left 355 papers for further screening following PRISMA methodology, as described later in Section 4.

The database search using Q2 yielded in total 533 records. Using Rayyan screening tool [23], 165 duplicates were automatically detected and removed. Of the remaining 368 records, 33 additional records were excluded after manual inspection: 14 records corresponding to entire conference proceeding volumes, 2 books, 3 book chapters, and 14 duplicates or superseded versions (e.g., earlier conference papers subsequently extended into journal publications). This left 335 papers for further screening following PRISMA methodology, as described later in Section 5.

As Q1 and Q2 target distinct literatures and require different domain-specific inclusion criteria, the eligibility criteria applied to each resulting corpus are described separately in Sections 4 and 5, together with the included studies. This separation was chosen to make each thematic strand easier to follow, allowing readers to trace the eligibility criteria and results for compositional generalization and for Croatian/South Slavic ASR independently before they are brought together in the synthesis

4. COMPOSITIONAL GENERALIZATION IN NLP

The 355 papers that remained after identification stage were subjected to title and abstract screening, as shown in Figure 4.1. Following this stage, a total of 264 papers were excluded based on predefined eligibility criteria, falling into two broad reasons for exclusion. The first, and conceptually more basic, reason was term ambiguity. We excluded 31 papers because they were retrieved through keyword matches in the title, abstract, or keywords, even though the matched terms were used in an unrelated technical or everyday sense rather than in the context of compositional generalization. Unlike the wrong-domain papers discussed below, these studies were not concerned with language or computation. Instead, they were false positives caused by lexical ambiguity in database searches, for example, the terms "systematicity", "generalization", or "compositional" being used in unrelated contexts.

The largest sub-group consisted of papers addressing semantic compositionality as a linguistic or distributional-semantics phenomenon rather than as a benchmark-evaluated model property (n = 69). These records surfaced because they used the terms "compositional" or "compositionality" in their theoretical or lexical-semantic sense, without testing or benchmarking systematic generalization in a trained model. The second-largest sub-group concerned figurative language and word-sense disambiguation (n = 34), where "compositional" described the interaction between literal and figurative meaning rather than a model's capacity to generalize to novel combinations. A further cluster spanned general linguistics, humanities, and morphology (n = 20), none of which involved a computational or model-evaluation component. Sentiment analysis papers (n = 18) formed the next notable subgroup, in which "compositional" referred to how sentiment is aggregated across sub-expressions rather than to systematic generalization. General NLP/ML and text-generation papers (n = 17) covered a broad range of applications that referenced compositional structure only incidentally. Vision and robotics papers (n = 15) used "compositional" in the sense of compositional scene understanding or task planning, which is a related but distinct notion from the linguistic compositional generalization addressed by this review.

Information retrieval (n = 10) and a semantics-related cluster (semantics, semantic analysis, semantic reasoning, and natural language inference; n = 10) each matched the search terms without directly evaluating compositional generalization. Knowledge representation and explainable AI (XAI) papers (n = 8) and machine translation papers (n = 6) that did not specifically test compositional generalization splits were also excluded. Several smaller clusters together accounted for the remainder: lexical resources, quantum computing, neuroscience and cognitive science, medical and health sciences, and parsing theory each contributed four to five papers, alongside two papers on education. One further paper did not fit any clear category and was excluded as wrong domain without subclassification.

The remaining 91 papers were retained for full-text review. In the second phase, those 91 full-text manuscripts were screened against eligibility criteria, resulting in the exclusion of a further 15 publications – papers that did not meet the inclusion criteria upon full-text review and records for which the full text could not be retrieved. The remaining 76 publications were retained for inclusion in the review and organized in the following sections.

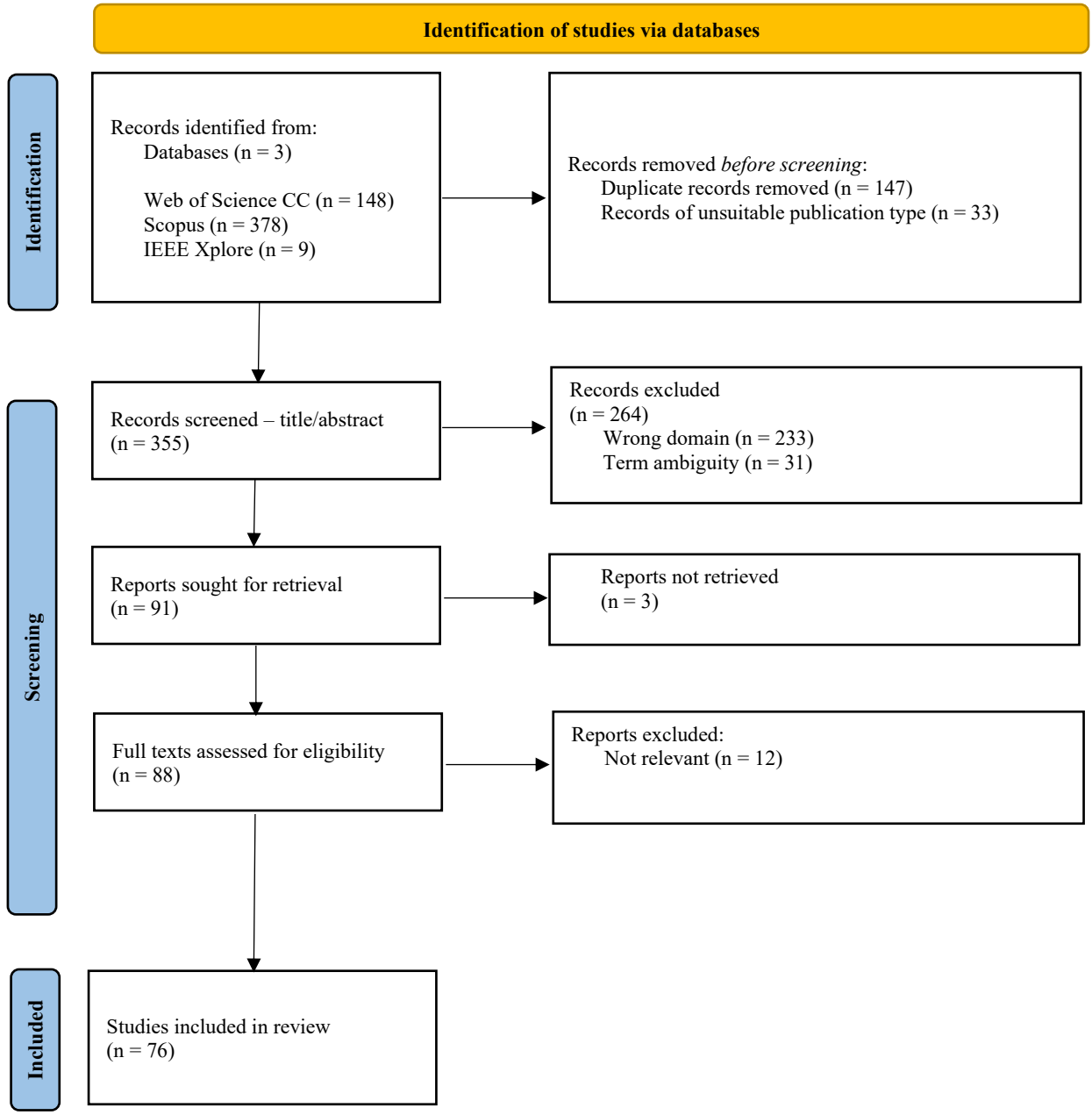

Figure 4.1. PRISMA flow diagram of the publication identification and screening process for Q1.

4.1. Benchmarks and diagnostic methods

Because ordinary held-out accuracy cannot detect compositional failure, the field has built specialized benchmarks that engineer a distributional gap between training and test compositions. The most influential is COGS, introduced in 2020 by Kim and Linzen [2]: a semantic parsing dataset whose evaluation split contains systematic gaps, namely novel combinations of familiar syntax and lexical items, resolvable only through genuine compositional generalization. Their original experiments found in-distribution accuracy near-perfect at 96 to 99%, while generalization accuracy dropped to just 16 to 35%. Petit et al. [24] later showed that adding a supertagging step with valency information substantially improves

structural generalization on COGS, confirming the structural constraints measured via exact-match accuracy are important for generalization in semantic parsing.

More recently, COGS itself has faced methodological criticism. Jabbar et al. [25] argue that a subset of its hardest generalization cases are not fair tests of compositional ability, since the correct output for these cases is underdetermined by, or even contradicted by, the training data. They show that resolving this ambiguity through targeted pretraining substantially reduces the generalization gap without changing the architecture model. This suggests that the widely cited 16–35% COGS generalization accuracy might reflect both genuine compositional deficits as well as an artifact of an underspecified evaluation task.

A second key benchmark is the Compositional Freebase Questions (CFQ) benchmark of Keysers et al. [8]. It creates splits by maximizing "compound divergence", which is dissimilarity between train/test compound distributions, while keeping "atom divergence" constant. This helps isolate novel-compound difficulty from vocabulary shift. Across three architectures (LSTM with attention, Transformer, Universal Transformer), accuracy exceeds 95% on random splits but falls below 20% on maximum-divergence splits despite comparable training set size and atom distribution, revealing a strong negative correlation between compound divergence and accuracy. The authors take this as evidence that models capture superficial rather than compositional structure, a pattern they replicate by applying the same construction method to SCAN. Its companion study, *-CFQ [26], examines how compositional generalization changes with training size under fixed computational budgets. It finds that the challenge persists at all training sizes and that increasing the scope of natural language leads to consistently higher error rates, which additional data only partly reduces.

Several additional benchmarks investigate complementary aspects of compositional generalization. gSCAN [27] extends SCAN to a grounded, instruction-following setting situated in a grid world, testing systematic generalization to novel adjective-object and verb-adverb combinations (e.g., interpreting an unseen color-object pairing or an unfamiliar manner adverb applied to a known verb). SyGNS [28] tests whether models can map novel combinations of logical expressions (quantifiers, negation) to scoped meaning representations, whereas CTL++ [29] inherits the original CTL's [30] use of unary symbolic function composition but shifts the target phenomenon: whereas CTL tests length generalization (productivity to longer, unseen compositions), CTL++ isolates systematicity by testing generalization to unseen combinations of already-known functions, holding composition length fixed.

A closely related diagnostic effort works in the opposite generation direction: Wang et al. [31] build an end-to-end neural generator that maps formal Discourse Representation Structures (DRSs) back into English sentences, introducing five challenge sets (tense, grammatical number, polarity, named entities, quantities) and a semantic-sensitive metric, ROSE, to assess systematicity and generalization to unseen inputs. Evaluation shows that negation is the most difficult phenomenon for DRS generation, followed by tense, while models perform relatively well on unseen quantities and names.

Alongside these task-specific benchmarks, a parallel line of research has focused on evaluation methodology itself. Goodwin et al. [32] define systematicity linguistically and construct probes showing neural models often generalize non-systematically even with strong held-out accuracy. Manino et al. [33] propose metamorphic testing relations targeting systematicity, compositionality, and transitivity directly, rather than robustness alone. Sun et al. [9] take a skeptical cross-benchmark view, showing that model rankings vary across datasets and splitting strategies. A later replication confirmed that the best-performing model on SCAN does not necessarily perform best on other datasets, raising concerns about the generalizability of results from any single compositional generalization benchmark. This concern is also supported by Valvoda et al. [34], who generate large sets of artificial languages using finite-state transducers and find that neural models either learn a compositional relation fully or fail completely. In their analysis, learnability depends mainly on transition coverage (around 400 training examples per transition), rather than model architecture. Notably, SCAN is an exception to this pattern, suggesting that no single measure of formal complexity can fully explain benchmark difficulty and that findings from individual hand-crafted datasets may not generalize. Addressing this measurement problem directly, Wold et al. [35] link task difficulty to the entropy of the training distribution, showing that systematic-generalization performance across RNN, CNN, and Transformer architectures scales predictably with this entropy, even without built-in compositional priors. This suggests that some of the cross-benchmark differences reported by Sun et al. and Valvoda et al. may be due to differences in distributional entropy rather than true differences in compositional ability.

Further diagnostic studies probe model internals rather than input–output behavior alone. Sachan et al. [36] were among first to ask whether text classifiers' predictions are grounded in genuinely compositional processing, a question Dasgupta et al. [37] later formalized into a diagnostic framework for measuring abstract, composable structure in learned sentence representations. Guo et al.'s [38] extend this work by proposing a two-step method to measure additive compositionality in word, sentence, and knowledge-graph embeddings. They find that static, transformer-based, and graph embeddings all show measurable but imperfect additive structure. Compositionality generally increases during training and peaks in intermediate transformer layers before declining near the output layer. The method also identifies cases where linear composition fails across all model types.

Fu and Frank [39] extend this diagnostic benchmarking into graph-based commonsense reasoning, using concept sets organized as reasoning graphs rather than sequences or trees. Yao and Koller [40] show that seq2seq difficulty generalizing to unseen structures is not limited to semantic parsing but also occurs in syntactic parsing and text-to-text tasks, and that this limitation can often be overcome by neurosymbolic models with linguistic knowledge built in. Galke et al. [41] provide further evidence that small RNNs trained from scratch, large pretrained language models used via in-context learning, and human learners all show more systematic generalization when trained on more structured languages, with similar effects on memorization. Together, these studies form an important empirical basis for the field while showing that conclusions about compositionality are sensitive to benchmark design.

4.2. Architectural approaches

One major strand treats CG as fundamentally an architectural problem, arguing that standard sequence-to-sequence models lack the inductive biases needed to generalize systematically. Gordon et al.'s [42] foundational proposal reframes compositionality as group-equivariance, building equivariant sequence-to-sequence architectures that significantly outperform regular seq2seq and convolutional baselines on SCAN. This framework was extended by Basu et al. in two papers: Equi-Tuning [43] converts non-equivariant pretrained models into group-equivariant ones with minimal representational loss, outperforming both fine-tuning with and without data augmentation, while their companion paper on efficient equivariant transfer learning [44] finds that naïve group-averaging performs poorly on equivariant zero-shot tasks despite good fine-tuning results, motivating a learned weighting scheme that outperforms it on both fronts. Ontañón et al. [45] instead search the design space of standard Transformers themselves, showing several configuration choices substantially affect compositional generalization. Their best standard seq2seq transformer configuration reaches 47.5% accuracy on the COGS generalization set, while reformulating the task using an intermediate sequence tagging representation yields 78.4% accuracy. Both significantly outperform the 35% baseline achieved by the best previously reported model [2], and raise average sequence-level accuracy from 13.7% to as high as 52.7% across all twelve tasks (with their specialized intermediate representations raising CFQ semantic parsing accuracy to 55.5%).

A second cluster builds explicit hierarchical or tree structure into models. Patel and Flanigan's Treeformers [46] induce nested hierarchical structure over Transformer inputs, addressing the architecture's lack of bias toward hierarchical language structures. Wang et al.'s Structured Reordering model [47] learns segment-to-segment alignments as discrete latent variables by combining reordering with monotonic alignment. The approach is inspired by traditional grammar formalisms, which use similar alignments to support systematic generalization. Lindemann et al.'s SIP method [48] adds a structural inductive bias to seq2seq models via simulation-based training, improving systematic generalization in Transformers, including generalization to longer inputs. Related tree- and graph-structured proposals include Guo et al.'s hierarchical poset decoding [49], which formalizes language understanding as predicting a partially ordered set (poset) represented as a DAG to enforce partial permutation invariance, and Gai et al.'s Grounded Graph Decoding [50], which preserves input syntax context through a grounded attention mechanism while decoding target queries as graphs, solving CFQ’s MCD1 split with 98% accuracy and improving generalization over flat-embedding baselines.

A third cluster targets modular and relational computation. Bergen et al.'s Edge Transformer [51] associates vector states with edges (pairs of input tokens) rather than nodes, drawing on both Transformers and rule-based symbolic AI by implementing a triangular attention mechanism inspired by unification in logic programming to support systematic relational reasoning. It reaches state-of-the-art accuracy of 87.4% on COGS, and exhibits an 18 to 41 percentage-point exact-match advantage over baselines on the more challenging MCD2 and MCD3 splits of CFQ dependency parsing. Russin et al. [52] factorize the alignment

and translation components of a seq2seq model to reflect the cognitive-science distinction between syntactic and semantic processing systems, allowing known grammatical rules to apply systematically to novel words. Chakravarthy et al. [53] investigate the mechanistic origin of systematicity within attention itself, showing that a separate "role" stream, which controls attention queries and keys while a filler stream determines values, causes systematic generalization to improve when abstract grammatical role labels guide that stream.

Extending this relational perspective beyond text-based tasks, Riveland and Pouget [54] show that RNNs can generalize zero-shot to novel sensorimotor tasks when guided by natural-language instructions encoded with a pretrained language model. This generalization relies on a shared compositional structure between the instruction embeddings and task representations, demonstrating that language-conditioned models can achieve systematic generalization even when the output is non-linguistic. Earlier neural parsing and recurrent-network studies, however, found more limited systematicity: Lane and Henderson [55] explored architectural features, namely generalization across syntactic constituents and bounded resource effects, that improve neural parsing effectiveness. In sentence processing, Frank [56] found that simple recurrent networks (SRNs) exhibit limited "weak combinatorial systematicity" due to overfitting in larger networks, where an excess of trainable parameters causes the model to memorize training sequences. Crucially, Frank demonstrated that Echo State Networks (ESNs) overcome this limitation; by leaving recurrent reservoir connections untrained, ESNs expand memory capacity without the overfitting bottleneck, yielding superior generalization. Exploring alternative architectures, Farkaš and Crocker [57] showed that a recursive self-organizing network paired with a prediction module can match the systematicity of ESNs while providing more robust, lesion-resistant context representations than SRNs. Recently, Zhang et al. [58] showed that complexity-control strategies, such as parameter initialization scale and weight decay, shape whether modern Transformers learn generalizable primitive-level rules versus memorized shortcuts on reasoning tasks. They found that smaller initialization scales promote a lower-complexity neuron condensation effect that reduces matrix stable rank, encouraging out-of-distribution generalization validated across both compositional image-generation and natural language processing benchmarks.

4.3. Training and data-augmentation approaches

A complementary strand holds that CG can be improved not by changing architecture but by changing training data and procedure. Li et al. [59] use two separate representations of the input sentence, one generating attention maps, one mapping attended input to output, to make prior compositional knowledge directly usable during training. Qiu et al. contribute two studies: one [60] evaluates encoder-decoder models up to 11B parameters and decoder-only models up to 540B parameters, finding that fine-tuning shows flat or negative scaling curve on compositional generalization while in-context learning improves with the scale but still underperforms much smaller fine-tuned models. The other proposes the Compositional Structure Learner [61], a more powerful data-recombination method than prior task-specific augmentation techniques, achieving 99.5% generalization accuracy on COGS and improvements over SCAN, GeoQuery, SMCalFlow-CS as well.

Data augmentation recurs throughout this literature. Zhang et al.'s TreeMix [62] proposes constituency-based data augmentation that leverages compositional sub-part structure, unlike prior techniques such as noise injection or back-translation, and outperforms these state-of-the-art augmentation methods on standard benchmarks. Fang et al. [63] explore using ChatGPT itself as a data-generation engine to improve compositional generalization in open intent detection, finding it outperforms back-translation, synonym substitution, and noise injection in both accuracy and F1-score across three benchmarks. Wei et al. [64] examine two strategies for compositional generalization in kinship prediction from stories: data augmentation, which improves generalization accuracy by around 20% on average over baseline, and intermediate kinship-graph representations, which counter-intuitively deteriorate generalization by around 50% relative to the data-augmentation-only model.

Curriculum and continual learning form a related cluster. Owoeye et al. [65] propose curriculum compositional continual learning for neural machine translation, building on a prior syntax and semantics disentangling approach and showing curriculum ordering further improves generalization. Fu and Frank [66] introduce the Continual Compositional Generalization in Inference (C2Gen NLI) challenge, arguing that standard evaluations unrealistically assume full advance access to primitive inferences, unlike humans who acquire such knowledge incrementally. They find that models fail to compositionally generalize under this continual setup, that standard continual learning methods such as Experience Replay and A-GEM reduce forgetting but still underperform offline training, and that ordering primitives from easy to hard while respecting task dependencies improves compositional generalization.

Two further contributions carry particular conceptual weight. Liu et al.'s LeAR model [67] targets algebraic recombination directly, which refers to the recursive recombination of structured expressions, and argues that prior augmentation methods recombine only lexical units, a necessary but insufficient part of full algebraic recombination. On two realistic compositional generalization benchmarks, LeAR improves accuracy from 67.3% to 90.9% on CFQ and from 35.0% to 97.7% on COGS relative to baseline models.

Akyürek and Andreas [68] contribute two related papers reframing training around lexical structure: their lexicon-learning work shows many systematic-generalization failures stem from an inability to disentangle lexical content from syntactic and semantic structure, while their subsequent LexSym paper [69] generalizes this into a domain-general, model-agnostic formulation of compositionality as a constraint on distributional symmetries rather than architecture, matching or surpassing task-specific baselines on COGS, SCAN, ALCHEMY, and CLEVR-COGENT, including a roughly 33% relative error reduction on CLEVR-COGENT. Kim et al. [70] propose annotating training data with explicit parse structure to guide models toward the hierarchical decomposition humans use when parsing novel sentences. In machine translation specifically, Shen et al. [71] propose consistency regularization encouraging models to interpret symbols consistently across contexts, and on CoGnition benchmark their method reduces instance-level compound translation error rate from 30.8% to 20.2% and aggregate-level error rate from 63.5% to 48.3%. When combined with LLM-based data augmentation, it surpasses a fine-tuned LLaMA2-13B model while using only about 0.2% of its parameters.

Chaabouni et al. [72] temper enthusiasm for synthetic-benchmark gains, finding that SCAN-capable Transformer modifications fail to transfer to resource-rich translation but yield up to 13.1% BLEU gains in low-resource settings and a 14% improvement on a novel distribution-shifted English-French task.

4.4. Compositional generalization in semantic parsing, text-to-SQL, and knowledge-based question answering

Semantic parsing has served as a natural application domain for CG research, since mapping language to structured meaning makes compositional structure explicit. Damonte and Monti [73] unify training across five heterogeneous semantic parsing datasets via multi-task learning, showing that a fully shared architecture matches or exceeds single-task baselines while cutting parameters by 68%, and that this multi-task approach also produces better compositional generalization than training separate single-task models. Shaw et al. [74] ask whether a single semantic parsing approach can handle both natural language variation and out-of-distribution compositional generalization simultaneously, since architectures optimized for compositional bias are typically validated only on synthetic data underrepresenting authentic variation. Their hybrid NQG-T5 model outperforms existing approaches across several compositional generalization challenges on non-synthetic data, though the authors note that no existing method performs well across the full range of evaluations, leaving the underlying problem unsolved. This synthetic/natural distinction recurs directly in machine translation: Dankers et al. [75] re-instantiate three compositionality tests from the literature for neural machine translation, finding that models trained on more data are, counter-intuitively, more compositional, and that some apparently non-compositional behaviors reflect genuine variation in natural data rather than model failure, reinforcing the need for non-synthetic evaluation across NLP tasks generally, not semantic parsing alone.

Text-to-SQL parsing offers a particularly active sub-area given SQL's clausal structure, and several works target this structure directly. Gan et al. [76] construct a clause-level compositional test set (Spider-CG) by splitting sentences into sub-sentences, finding that models degrade significantly even on sub-sentences seen during training, while training on their segmented data improves generalization beyond prior word-level augmentation. Building on this clause-level focus, Arora et al. [77] improve LLM-based cross-compositional generalization through offline selection of a minimal exemplar set with complete SQL clause and operator coverage, avoiding costly inference-time retrieval and outperforming generic prompting on KaggleDBQA, while Parthasarathi et al. [78] extend the same compositional challenge into multi-turn, conversational text-to-SQL, showing that multi-task training with reranking yields modest gains over a state-of-the-art baseline on CoSQL but that generalization to unseen parse trees in context-dependent turns remains unresolved.

Knowledge-based question answering (KBQA) presents a related compositional challenge, requiring composition of atomic facts into logical forms. Chen et al.'s CERG-KBQA framework [79] addresses coverage and generalization limitations, especially in compositional and zero-shot situations, through classification, enumeration, ranking, and generation strategies. A related decomposition-based strategy is proposed by Huang et al. [80], whose Question

Decomposition Tree (QDT) represents complex questions as a tree structure spanning multiple compositionality types simultaneously, rather than the single compositionality type assumed by prior decomposition methods; their QDTQA system improves an existing KBQA system by 11% and sets a new state of the art on LC-QuAD 1.0. Gai et al.'s Grounded Graph Decoding [50], introduced in Section 4.3, similarly targets question answering specifically, preserving syntactic context to improve generalization to longer or structurally novel queries.

Across this section, the general architectural and training innovations are adapted to structured tasks, while domain-specific properties such as SQL clause structure or dialogue context introduce their own compositional challenges beyond purely synthetic benchmarks.

4.5. Compositional generalization in LLMs

The rise of LLMs reframes the CG question: rather than engineering purpose-built architectures, researchers increasingly ask whether scale and in-context learning alone produce systematic behavior.

Hosseini et al. [81] find that a compositional generalization gap between in-distribution and out-of-distribution performance persists under in-context learning in semantic parsing across CFQ, SCAN, and GeoQuery, though the relative size of this gap decreases as models are scaled across four model families (OPT, BLOOM, CodeGen, and Codex), suggesting scale helps but does not eliminate deficit. Han and Pado [82] examine whether meta-training a causal Transformer to perform in-context learning, by exposing it to randomized trajectories of training examples with shuffled labels to suppress memorization, can provide an inductive bias that improves compositional generalization, treating in-context learning as an active training mechanism rather than merely a capability to evaluate. Qiu et al.'s scaling study [60], discussed in Section 4.3, is directly relevant here too, probing whether increasing model size improves compositional generalization now that scale dominates most other capability gains.

Mathematical and logical reasoning have become a favored proving ground, since multi-step reasoning inherently requires compositional combination of simpler inferential steps. Zhao et al.'s MathTrap [83] dataset introduces logical traps into MATH and GSM8K problems specifically to expose LLMs' compositional deficiencies in mathematical reasoning; accuracy on trap problems drops to roughly half of original-problem accuracy or lower across tested models, and even OpenAI's o1 model, whose human-like slow thinking improves compositionality (achieving 67.7% accuracy on traps), still falls short of the 83.8% accuracy achieved by human participants on the same trap problems (which rises to 95.1% with notice). Zhang et al. [84] examine whether human-guided tool manipulation can compensate when LLMs' internal reasoning proves insufficient for compositional tasks. Their proposed framework achieves state-of-the-art results on two compositional generalization benchmarks and outperforms existing methods by 70% on the hardest test split, with one specific improvement raising accuracy from 27.3% to 95.8%. Petruzzellis et al. [85] systematically benchmark GPT-4 on algorithmic problems across prompting strategies, finding that while advanced prompting techniques, especially self-consistency, let GPT-4 outperform GPT-3.5 and a specialized Neural Data Router baseline, no model or prompting method fully solves

deeply nested formulas, with accuracy degrading smoothly as nesting depth and operand count increase.

Testing a related but distinct reasoning setting, Lee et al. [86] evaluate LLMs on the Abstraction and Reasoning Corpus (ARC) through the lens of the Language of Thought Hypothesis, decomposing performance into logical coherence, compositionality, and productivity. They find current LLMs fall substantially short of human performance on all three components, with compositionality failures characterized by a weak ability to programmatically decompose tasks and combine step-by-step Domain-Specific Language (DSL) functions (where LLMs select correct sequences only 3% to 14% of the time, compared to 86% for humans). Furthermore, their logical coherence experiments show that LLM generalization fails entirely when tasks require integrating multiple prior knowledge domains rather than just a single simple concept.

A substantial body of work benchmarks and diagnoses LLM compositional behavior specifically. Sakai et al.'s Ordered CommonGen [87] jointly evaluates compositional generalization and instruction-following by requiring sentence composition that respects a specified concept order. Across 36 LLMs, even the best-performing model achieves only about 75% ordered coverage, and models often default to memorized concept-order patterns regardless of instructions. Schmidt et al.'s CompoST [88] benchmark probes the systematicity of LLMs' question-to-SPARQL interpretation in a QALD setting, finding that macro F1 degrades from 0.45 to 0.26 to 0.09 as test questions deviate further from familiar patterns, with F1 never exceeding 0.57 even on the least complex subset when all necessary information is provided. Kumon and Yanaka [89] explore Transformer internals directly, identifying a subnetwork with better generalization performance than the full model. However, they find that this subnetwork relies on a non-compositional algorithm acquired early in training alongside genuine syntactic features, complicating claims of purely compositional processing. Xu et al.'s MC^2 framework [90] proposes a minimum-coverage, dataset-agnostic approach to LLM compositional generalization in semantic parsing. The authors first show that under this minimum-coverage constraint current LLMs cannot achieve good generalization generically without dataset-specific engineering, and demonstrate that MC^2 effectively narrows this gap across multiple semantic parsing datasets. Zhang et al.'s complexity-control study [58], introduced in Section 4.3, is equally relevant here, since it investigates the internal mechanisms determining whether Transformers learn generalizable rules versus memorized shortcuts.

Collectively, this cluster suggests that scale alone has not resolved the systematic generalization deficits identified in earlier architectures — the problem appears to have migrated to new models rather than been solved.

4.6. Morphology, multilingual settings, and the path toward speech

While most CG research has been developed and evaluated on English or English-like synthetic languages, a small but critical cluster of studies begins to interrogate how compositional generalization behaves in morphologically richer and typologically diverse languages — the setting most directly relevant to this thesis. This cluster is small precisely

because it is underexplored, and its scope is itself diagnostic of the research gap this thesis addresses.

Vrabcová and Sojka [91] demonstrate how language-specific morphological and grammatical properties disrupt compositional behavior in ways English-centric benchmarks fail to capture. Focusing on Czech, a morphologically rich Slavic language, they show that verbal negation, expressed via short, prefixed “ne-“ morphemes that do not induce a discrete distributional change in learned representations, causes systematic failures in natural language inference. By extending the Czech factuality dataset CsFEVER with negated hypotheses to construct the CsNoFEVER corpus, they find that publicly available language models experience a drastic performance drop on negated hypotheses compared to positive baselines. The authors highlight that syntactic negation in a fusional language like Czech often requires modifying multiple words beside the verb, contrasting with the separate lexical items in English. This demonstrates that compositional phenomena trivial in English can become considerably harder when morphological marking carries the relevant compositional information, which is a finding this work argues is of critical consequence when these morphological structures are subjected to acoustic and speech processing conditions.

Moisio et al. contribute two closely related papers that apply the distribution-based compositionality assessment (DBCA) framework to neural machine translation (NMT). Their first paper [92] applies the DBCA framework to evaluate morphological generalization, demonstrating that smaller subword (BPE) vocabularies significantly improve an NMT system's capacity to generalize to unseen inflected word forms. Their second paper [93], submitted to the GenBench collaborative benchmarking task, extends DBCA to natural, syntactic dependency relations in the Europarl corpus, isolating compositional generalization performance in a real-world translation task rather than a synthetic benchmark. These two papers establish DBCA as a powerful, automated framework for partitioning natural corpora to target distinct structural levels (morphological features and syntactic dependency relations) as first-class axes of compositional generalization.

Complementing this translation-focused work, Ismayilzada et al. [22] evaluate morphological compositional generalization directly in LLMs, probing whether progress in text generation and understanding extends to morphology. In typologically rich agglutinative languages, humans systematically and productively inflect novel words based on learned rules. Testing instruction-tuned multilingual models (including GPT-4 and Gemini) on the agglutinative languages Turkish and Finnish, the authors find that LLMs struggle particularly with novel word roots, with performance declining sharply as morphological complexity increases. Furthermore, while models identify individual morphological combinations better than chance, their performance lacks systematicity, resulting in significant accuracy gaps compared to human benchmarks. This highlights morphology as a distinct and still unresolved challenge for compositional systematicity.

The multilingual dimension is addressed most directly by Wang and Hershcovich [94], who note that because most CG research is conducted in English, language-specific compositional differences and cross-lingual systematicity are heavily underexplored. They

demonstrate that prior benchmarks translated via neural machine translation introduce critical semantic distortion, prompting them to construct a rule-based translation of the MCWQ dataset from English into Chinese and Japanese. Even with this more robust benchmark, they find that the distribution of compositions still suffers due to linguistic divergences, and multilingual models continue to struggle with cross-lingual compositional generalization.

Finally, and most directly relevant to our work, Ray et al. [21] present the only paper in this literature investigating compositional generalization in spoken language understanding (SLU). Their study identifies two types of compositionality relevant to SLU, novel slot combinations and length generalization, finding that state-of-the-art SLU models frequently learn spurious slot-to-slot correlations during training that degrade performance on unseen slot distributions. To mitigate this, they propose a compositional SLU model trained with a specialized attention-decoupling loss and a "paired training" data augmentation technique, outperforming a state-of-the-art BERT SLU baseline by up to 5% in slot-tagging F1 score on compositional splits of ATIS and SNIPS. While this work establishes a vital baseline for systematicity in SLU, its scope remains structurally limited: it operates purely on clean, textual transcripts of English queries. Consequently, it does not address typologically rich morphological systems (such as agglutination or fusional inflections), nor does it engage with the acoustic channel, phonetic ambiguities, or ASR errors that are central to the spoken domain.

Taken together, these six studies reveal a clear research gap at the intersection of CG, morphology, and speech processing. While morphological CG has been studied in text-based environments, and multilingual benchmarks have examined the cross-lingual dimension, only one study has explored CG in the spoken language domain. However, this SLU study uses clean English text transcripts and therefore does not address either morphological complexity or the acoustic channel. Similarly, the existing morphology-focused studies, including work on Czech verbal negation, machine translation, and LLM inflection, are entirely text-based and do not consider speech or ASR. This leaves an important gap in understanding how compositional generalization interacts with morphology in speech. In MRLs, acoustic and morphological ambiguities can interact: a single speech recognition error may change or obscure grammatical features such as case, number, or aspect. Despite this challenge, existing CG research has not yet developed benchmarks, models, or training methods that address the interaction between acoustic and morphological variation.

5. CROATIAN AND SOUTH SLAVIC ASR

The 335 papers that remained after identification stage were subjected to title and abstract screening, as shown in Figure 5.1. Following this stage, a total of 169 papers were excluded based on predefined eligibility criteria, with the primary reason for exclusion being wrong domain, across several distinct sub-categories.

The largest sub-group consisted of linguistics, humanities, and pedagogical papers (n=84). These records surfaced due to the South Slavic language keywords in the search query but addressed topics such as morphosyntactic analysis, discourse markers, code-switching, dialect documentation, lexicographic studies, literary analysis and cultural history without any computational speech recognition component. Next notable sub-group of wrong-domain papers concerned text-to-speech (TTS) and speech synthesis systems (n=25). Although TTS and ASR share acoustic modelling infrastructure, these papers did not report speech recognition results. Medical and health sciences papers formed the third-largest subgroup (n=17), predominantly epidemiological studies on cancer incidence and mortality in Serbia and Croatia that matched the geographic and language filters but have no relation to ASR.

Computational linguistics and NLP-only papers (n=13) described resources such as morphosyntactic taggers, dependency treebanks and text corpora which, while potentially useful as upstream tools in a speech pipeline, presented no speech recognition experiments or acoustic modelling. Next excluded sub-group of papers addressed dialogue systems (n=12), as ASR functioned as one pipeline component rather than the primary research contribution. Speaker recognition and classification papers (n=9) included tasks which rely on similar acoustic front-ends as ASR, however, they optimize for speaker identity rather than linguistic content, and their evaluation metrics (equal error rate, detection cost) differ fundamentally from ASR metrics. Four remaining papers fell into other wrong-domain categories, and five papers were excluded due to wrong language, having evaluated languages outside the scope of this review.

The remaining 166 papers were retained for full-text review. In the second phase, those 166 full-text manuscripts were screened against eligibility criteria, resulting in the exclusion of a further 82 publications – papers that did not meet the inclusion criteria upon full-text review, records for which the full text could not be retrieved, and studies available only in languages other than English or Croatian. The remaining 84 publications were retained for inclusion in the review and organized in the following sections. During the writing of the review, two additional references were incorporated to support the narrative (e.g., providing background on a referenced initiative or dataset), bringing the total number of publications discussed in this section to 86.

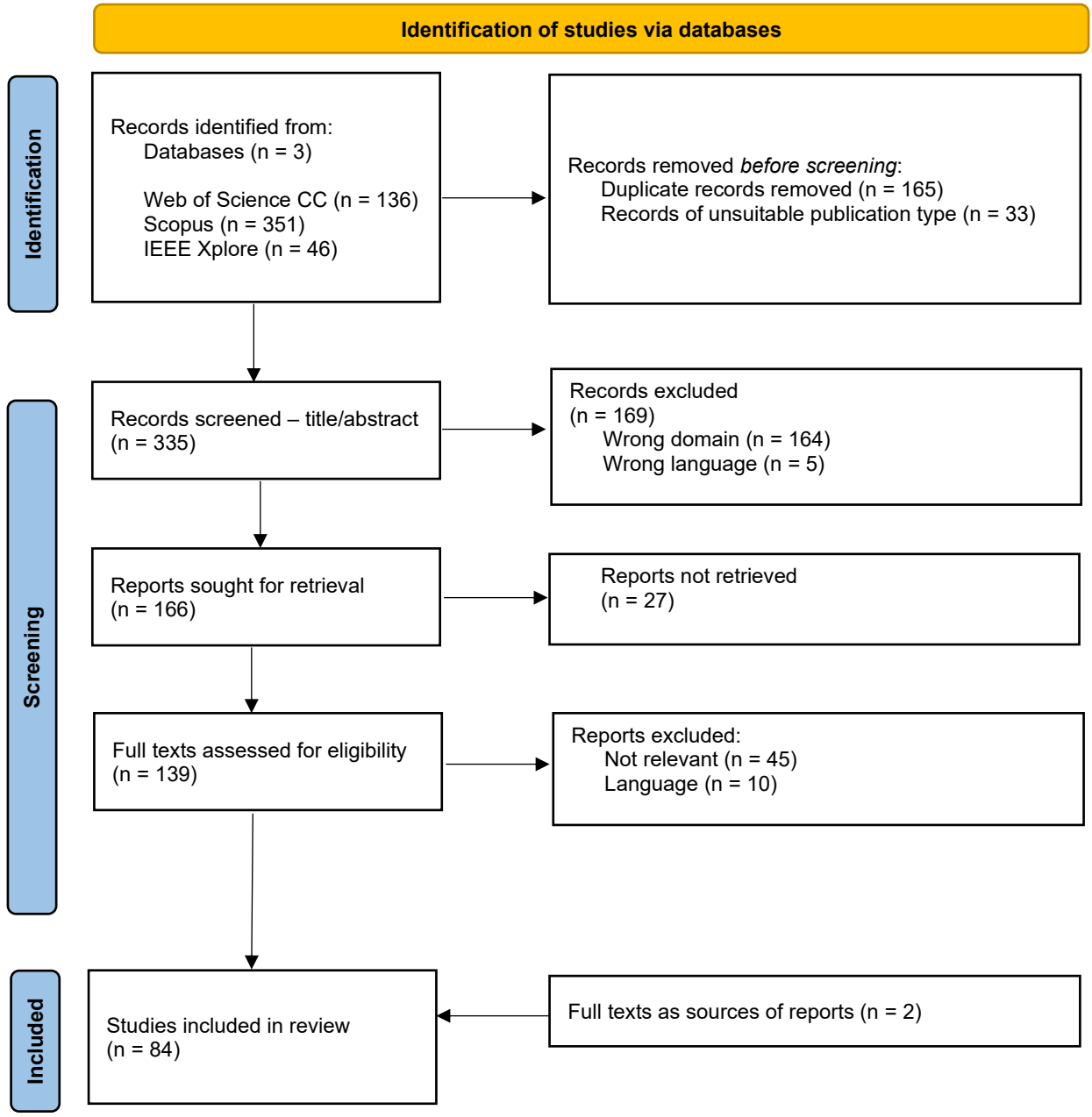

Figure 5.1. PRISMA flow diagram of the publication identification and screening process for Q2.

5.1. Speech corpora and language resources

Speech corpora and related language resources form the empirical foundation of ASR development for Croatian and other South Slavic languages. As summarized in Figure 5.2, the reviewed literature shows a noticeably uneven language distribution within this resource layer: most corpus and lexicon papers concern Slovenian, a smaller but still substantial group focuses on Croatian, and only a limited number describe Serbian resources directly. This pattern is not accidental, but follows the historical development of the field: Slovenian benefited early from large coordinated resource-building efforts such as SpeechDat(II), broadcast-news corpora, and later reference-corpus expansions, while Croatian and Serbian corpora were more often created in connection with specific applications, domains, or individual research groups. Taken

together, this uneven landscape illustrates why the availability of speech corpora has long been a practical bottleneck for South Slavic ASR development, particularly for Croatian and Serbian.

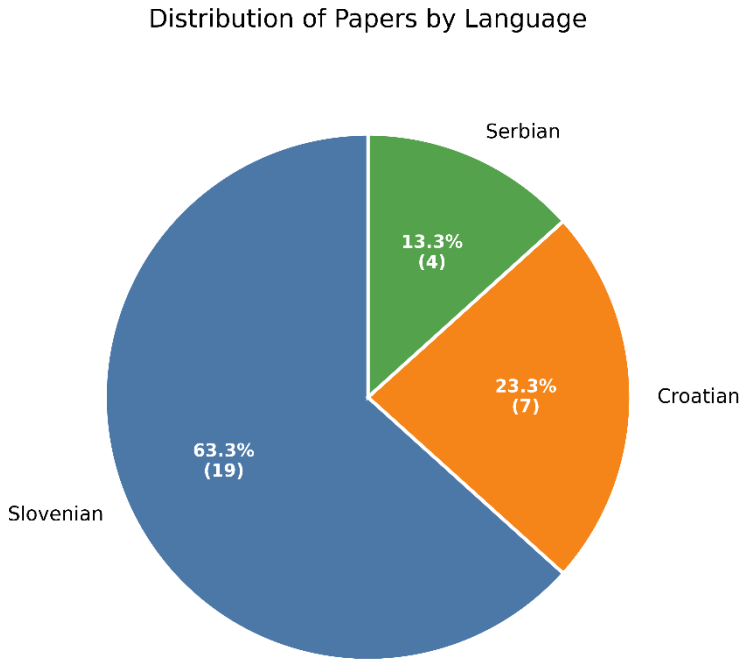

Figure 5.2. Language distribution of corpus-focused papers.

5.1.1. Large multilingual initiatives

The earliest efforts to address this scarcity were driven by large-scale European collaborative initiatives. The GlobalPhone project at the University of Karlsruhe, launched in 1995, assembled high-quality read-speech data for 15 languages including Croatian, making it one of the first multilingual corpora to include a South Slavic language in the context of Large Vocabulary Continuous Speech Recognition (LVCSR) research. The GlobalPhone collection comprises approximately 328 hours of transcribed speech from 1506 native speakers overall, of which the Croatian component accounts for about 16 hours of audio from 92 speakers and roughly 120,000 spoken words. Two papers in the reviewed corpus directly describe and evaluate the GlobalPhone database: the original system description [95] and the earlier language-independent LVCSR framework paper [96], which reports the first recognition experiments using GlobalPhone Croatian data.

Concurrently, the European SpeechDat(II) project [97] produced telephone-channel speech databases for a range of official European languages, including Slovenian. The Slovenian SpeechDat(II) fixed-network database (FDB-1000) contains recordings from 1,000 speakers, covering phonetically rich sentences, digits, natural numbers, spellings, dates, and spontaneous utterances. Within the reviewed literature, this resource surfaced through the study by Iskra et al. [98], which provides a detailed description of the Slovenian SpeechDat(II) corpus and reports recognition experiments across multiple task types including digit and natural

number recognition tasks of particular relevance to this review. Building on this infrastructure, Kačič et al. [99] discuss the practical issues in the design and collection of a large telephone speech corpus for Slovenian and introduce the SNABI corpus, which follows the SpeechDat(II) specifications while adapting text selection, speaker recruitment, and recording procedures to national constraints. Complementing the fixed-network corpora, the PoliDat database was later collected as a GSM/mobile-channel counterpart, with a similar prompt design but substantially different acoustic conditions. In the reviewed literature, PoliDat is introduced by Žgank et al. [100], evaluating it through recognition experiments that highlight the performance gap between fixed and mobile channels, providing an early benchmark for robustness studies in Slovenian ASR.

The COST278 pan-European broadcast news database, developed through a seven-institution European collaboration, extended multilingual resource coverage to include Slovenian alongside six other languages [101]. The corpus was designed as a shared audio-visual benchmark for broadcast news ASR and established a common evaluation framework across European languages. At the other end of the timeline, Mozilla Common Voice, described in the reviewed literature by Ardila et al. [102], brought crowdsourced open-access speech collection to multiple South Slavic languages including Slovenian, Croatian, and Serbian, primarily targeting ASR and related tasks such as keyword spotting and language identification.

5.1.2. Broadcast news, parliamentary, and large-scale spoken corpora

Within the Croatian research community, the most significant recent corpus contribution is ParlaSpeech-HR [103], the first large freely available Croatian ASR dataset at 1,816 hours, bootstrapped from parliamentary recordings and transcripts within the ParlaMint corpus [104]. The bootstrapping methodology relied on commercial ASR alignment followed by fine-tuning of a multilingual transformer, yielding word-level timing alignments and speaker metadata for 309 speakers, released under CC BY-SA 4.0 license. ParlaSpeech-HR is particularly notable as a template for large-scale, low-cost corpus creation in under-resourced Slavic languages.

For Slovenian, the largest and most impactful corpus development effort centered on broadcast speech. The Slovenian BNSI Broadcast News database [105], produced through a collaboration between the University of Maribor and RTV Slovenia, substantially expanded this coverage, providing 36 hours of transcribed television news recordings across 1,565 speakers from the period 1999–2003. The BNSI corpus became the standard evaluation benchmark for Slovenian LVCSR throughout the 2000s and is cited as the primary training resource in numerous papers. The SloParl corpus [106] further extended the domain coverage for Slovenian LVCSR development by contributing 100 hours of parliamentary speech from the period 2000–2005. In a subsequent effort, the SI TEDx-UM database [107] assembled 242 TEDx talks amounting to 54 hours of Slovenian presentational speech, with transcriptions generated automatically and evaluated against manually transcribed development subsets.

The Slovenian Gos 2 reference corpus, introduced by Verdonik et al. [108], represents the most recent large-scale update, extending the original GOS corpus to over 300 hours and

2.4 million words by integrating Gos VideoLectures and ARTUR corpus data. The corpus creation experience documented in the accompanying ARTUR paper by Verdonik et al. [109] offers a methodologically detailed account of the practical challenges in spoken language corpus development for a less-resourced language - specifically, copyright management, precise forced-alignment transcription, and demographic diversity - providing a replicable methodology applicable to Croatian and Serbian resource building.

5.1.3. Domain-Specific and Read-Speech Corpora

Alongside the large general-purpose corpora, several smaller domain-specific resources were created to support targeted ASR development. VEPRAD [110], a Croatian speech database of weather-forecast speech collected from Croatian Radio-Television (HRT) national broadcasts, was the primary training and evaluation resource for early Croatian ASR experiments.

The SINOD database [111] addresses a distinct but practically important gap: as the first Slovenian non-native speech database, it provides recordings from Russian and English native speakers and was designed explicitly as a supplement to the BNSI broadcast news corpus for non-native acoustic model adaptation.

For Serbian, specific acoustic conditions and speech styles have been addressed to improve robust speech recognition. The Whi-Spe whispered speech database [112], consisting of 50 isolated words pronounced in both normal and whispered modes by ten speakers, targets the specific challenge of register-dependent recognition performance and supports evaluation of ASR system robustness outside standard phonation conditions. Complementing these domain-specific resources, Ostrogonac et al. [113] developed a comprehensive set of speech resources for a Serbian LVCSR system. This initiative introduced a 200-hour corpus of segmented and transcribed broadcast audio alongside a pronunciation lexicon, significantly advancing baseline continuous speech recognition capabilities for the language.

In addition to these individual corpora, Delić et al. [114] provide a consolidated overview of speech and language resources for Serbian and kindred South Slavic languages developed over the last decade within joint projects of the Faculty of Technical Sciences in Novi Sad and the company AlfaNum, illustrating how much of the Serbian resource landscape has emerged from a single, tightly coupled research–industry collaboration.

5.1.4. Pronunciation Lexicons and Prosodic Annotation Resources

Pronunciation lexicons are a critical but often underemphasized component of HMM-based ASR pipelines, and their construction for highly inflected South Slavic languages is substantially more demanding than for morphologically simpler languages. The GOPOLIS database, described by Mihelič et al. [115], provides prosodically annotated Slovenian speech along with methods for automatic classification of prosodic events, establishing the first large-scale Slovenian prosodic annotation framework. The SIflex and SImlex lexicons, reported by Rojc et al. [116] and Rojc and Kačič [117], were developed semi-automatically from the Dictionary of Standard Slovenian to provide both phonetic and morphological lexicons for spoken language applications. In parallel, the LC-STAR II project [118] produced large

XML-encoded phonetic lexica for speech processing in ten European languages, including Slovenian, targeting automatic speech recognition and text-to-speech applications and further strengthening the Slovenian lexicon infrastructure within a multilingual EU framework. SI-PRON, described by Žganec Gros et al. [119], represents a later comprehensive machine-readable pronunciation lexicon for Slovenian; its coverage of inflected forms is particularly significant given the challenge of OOV rates in morphologically rich Slovenian.

Extending these morphological and phonetic layers into the syntactic domain, Dobrovoljc [120] introduced the Spoken Slovenian Treebank (SST). By providing manual annotations for lemmas, part-of-speech tags, and syntactic dependencies using the Universal Dependencies framework, this resource bridges the gap between spoken word recognition and higher-level grammatical understanding. While Slovenian benefits from this deep pipeline of phonetic, lexical, and syntactic annotations, no equivalent comprehensive pronunciation lexicon or spoken treebank for Croatian or Serbian was identified in the reviewed literature. This gap has direct consequences for downstream spoken language processing in these languages.

5.1.5. Specialized, Emotional, and Spontaneous Speech Corpora

Several corpora in the reviewed literature address specific speech registers or speaker populations that fall outside the standard read, broadcast, and parliamentary domains. The Croatian Emotional Speech corpus (KEG - *Korpus emocionalnog govora*), described by Dropuljić et al. [121], is the first emotional speech corpus for Croatian, comprising acted utterances annotated with five discrete emotional states across male and female speakers spanning multiple age groups. Its Slovenian counterpart, EmoLUKS [122], was built from 17 Slovenian radio dramas obtained from RTV Slovenia and annotated for emotional states across two annotation stages. While these emotional corpora are not primary ASR training resources, they serve as controlled evaluation conditions for studying how emotional speech degrades ASR performance.

Beyond acted emotion, several resources focus on spontaneous or richly annotated conversational speech. The Croatian Adult Spoken Language Corpus (HrAL), described by Kuvač Kraljević and Hržica [123], provides spontaneous conversational speech from 617 speakers across all Croatian counties and constitutes the most ecologically valid Croatian spoken language resource identified in the review; with over 250,000 tokens covering regional dialectal variation, speaking styles, and registers, it is better suited to spoken language research than to direct ASR system training. For Slovenian, the EVA multimodal corpus [124] offers deeply annotated multiparty conversational data with aligned speech and co-speech gestures, built primarily for pragmatic and conversational-intelligence research but potentially useful as a testbed for multimodal spoken-language understanding. On the Serbian side, the CRONUS corpus of spoken narratives [125] provides semi-spontaneous narrative speech with detailed discourse-level and construction-grammar annotation, again targeting discourse analysis rather than ASR training but supplying high-quality spoken data that could support specialized evaluation scenarios.

Finally, the Corpus of Spoken Istrovenetian/Fiuman and Croatian (C-ORAL-IC) [126], documents unscripted bilingual speech in the Croatian-Italian contact zone of the Istrian and Kvarner areas, and while its utility for standard Croatian ASR training is limited, it provides unique coverage of spoken Croatian in non-standard sociolinguistic conditions that may inform dialect robustness work.

Across all categories of speech corpora and language resources reviewed in this section, a consistent structural pattern emerges. First, resource development in South Slavic ASR is highly uneven, with Slovenian exhibiting a relatively mature and continuously evolving ecosystem of corpora, lexicons, and annotated datasets, while Croatian and Serbian remain more fragmented and application-driven in their resource construction. Second, most large-scale datasets originate from institutional or EU-funded initiatives, whereas Croatian and Serbian resources are more often produced through individual research efforts or task-specific projects, leading to a lack of long-term continuity and standardization. Third, there is a persistent domain imbalance, with broadcast news and parliamentary speech dominating across all languages, while spontaneous, conversational, and dialectal speech remains comparatively underrepresented, despite its importance for real-world ASR robustness. Finally, modern crowdsourced and bootstrapped approaches such as Common Voice and ParlaSpeech-HR indicate a clear shift toward scalable, low-cost corpus creation, suggesting a gradual mitigation of historical data scarcity.

5.2. Statistical ASR systems

Building upon the empirical foundation of the corpora and lexicons described in the previous section, the technical trajectory of South Slavic ASR has been primarily defined by the application and refinement of the Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) paradigm. This era, spanning from the mid-1990s to the early 2010s, is characterized by a transition from basic isolated word detectors to LVCSR systems. The "story" of this development is one of constant adaptation: researchers had to modify global standards like the HMM Toolkit (HTK) and Sphinx to accommodate the extreme morphological complexity and high inflection inherent to the Slavic branch of the Indo-European family.

5.2.1. Early telephone-based recognition and the SpeechDat foundation

The earliest experiments in South Slavic ASR focused on limited domains, most notably telephone-based tasks, especially isolated digits and connected natural numbers, which made them a practical starting point for early spoken dialogue systems in the region. Within this early phase, both Petek's Slovenian digit-recognition study [127] and Imperl's multilingual telephone-speech study [128] involving German and Slovene relied on the HMM Toolkit (HTK) framework, showing that constrained recognition tasks could be handled effectively. Petek's study (1995) reported 97.3% accuracy with gender-specific HMMs using 6-mixture components and Mel-Frequency Cepstral Coefficients (MFCC) features. This work was extended by Imperl (1999), who utilized the SpeechDat(II) database to develop multilingual connected-digit systems for Slovenian and German. A critical methodological finding from this period was that while unconstrained grammars sufficed for digits, the recognition of natural

numbers required explicit linguistic constraints to handle language-specific conjunctions (e.g., the Slovenian *"in"* in *dva in dvajset*), which significantly improved sentence recognition rates

The SpeechDat project gave this line of research a stronger empirical and methodological basis by introducing standardized fixed-telephone databases for numerous European languages, although in the South Slavic literature reviewed here it is represented predominantly through Slovene-focused resources and experiments. These databases, typically featuring recordings from 1,000 speakers per language, provided the foundation for the development of *reference recognizers* (RefRec). Implemented using Perl scripts and the HMM Toolkit (HTK), these baseline RefRec systems (described by [129]) became the standard benchmarking systems for subsequent South Slavic research.

While the literature predominantly features Slovenian-focused experiments using the Slovenian 1000 FDB SpeechDat(II) database, researchers utilized these resources for a wide array of technical advancements. For instance, Kotnik et al. [130] integrated an LDA-Toolkit into the RefRec to improve recognition through dimensionality reduction, while Vlaj et al. [131] used the Slovenian, German, and Spanish subsets to evaluate Voice Activity Detection (VAD) and frame attenuation strategies for improved noise robustness. Furthermore, Žgank and Kačič [132] leveraged the Slovenian SpeechDat corpus to explore the conversion of monophone models into grapheme-based acoustic models, which eliminated the need for complex phonetic vocabularies.

Methodologically, these resources allowed researchers to evolve from recognizing isolated words to handling connected digits and natural numbers. As demonstrated by Imperl, while simple, unconstrained grammars sufficed for producing strings of connected digits, the recognition of natural numbers required explicit linguistic constraints to handle language-specific conjunctions, such as the Slovenian "in" (e.g., *dva in dvajset* for 22) or the German "und". The integration of these simple grammatical rules proved vital, significantly improving sentence recognition rates in multilingual telephone dialog environments.

Foundational research also explored bilingual recognition for closely related variants. Žibert et al. [133] demonstrated a bilingual system for Slovenian and Croatian weather forecasts that utilized shared acoustic units across languages, finding that common phonetic rules could increase accuracy while reducing the number of total parameters. As the field diversified, alternative toolkits beyond HTK were explored; Gulić et al. [134] implemented a Croatian alphanumeric ASR using PocketSphinx, identifying that sound feature parameters like an 8 kHz sampling rate and a specific number of senones (approx. 200) were optimal for sound discrimination in the Croatian context.

However, once research shifted from these restricted domains toward LVCSR, the main challenge was no longer basic digit recognition, but the much harder Out-of-Vocabulary (OOV) problem caused by the rich inflectional morphology of languages such as Slovenian and Croatian.

5.2.2. The morphological challenge: sub-word modeling and lexical adaptation

The transition from restricted telephone-based tasks to Large Vocabulary Continuous Speech Recognition (LVCSR) for South Slavic languages exposed the limitations of traditional word-based language models. Slovenian, Croatian, and Serbian are characterized by rich morphology and high inflection, where grammatical categories such as case, number, gender, and tense are encoded through complex suffixation. Consequently, a single lemma can manifest in dozens of different word form (up to 18 for nouns and 100 for verbs) leading to a vocabulary growth rate approximately ten times faster than that of English, which results in an exceptionally high OOV rates [18]. To address this bottleneck, researchers introduced sub-word modeling, primarily through the decomposition of words into stems and endings, an approach applied to Slovenian by Maučec et al. [135]. Using data-driven algorithms like the longest-match principle and phonetic constraints (e.g., requiring endings to start with a vowel), they were able to reduce the Slovenian OOV rate from 18.2% to 4.5%.

Moreover, because sub-word units are shorter and acoustically more similar, they significantly increased the search-space complexity [18]. Novel search algorithms were developed to enforce the "legal" order of units, ensuring a stem was followed by a valid ending, which effectively limited the search space to a size comparable to word-based systems while increasing word accuracy by roughly 4.3% [135].

For highly inflected languages where vocabulary cannot be constrained to a static dictionary, the Hypothesis Driven Lexical Adaptation (HDLA) frameworks [136], [137] were proposed. HDLA uses a two-pass strategy: a first pass on a baseline dictionary identifies hypothesized words, which are then used to select morphologically or phonetically similar entries from a massive "fallback" lexicon for a second, more accurate recognition pass. This approach reduced OOV rates for Serbo-Croatian by up to 55%, leading to an absolute reduction in Word Error Rate (WER) of 4.1% [136].

5.2.3. Advancing acoustic modeling

As the complexity of South Slavic ASR systems grew, the data sparsity problem in acoustic modeling became a primary concern. Context-dependent triphone modeling was adopted to capture coarticulation, but the number of potential triphones in highly inflected languages often exceeds the training data's capacity, as triphones often represent only 12% to 22% of theoretically possible combinations [138]. To build robust models for these unseen triphones, state-tying procedures using phonetic decision trees became standard.

Martinčić-Ipšić et al. [139] developed sets of 83 language-specific phonetic rules (translated into 166 questions) to group acoustically similar HMM states. These questions explore the articulatory characteristics of a phoneme's neighbors, such as whether a left context is a nasal or a right context is a voiceless consonant. A major contribution to Croatian ASR was the Croatian Language Transcription (CLT) algorithm [140], which automated phonetic transcription by applying rules of phonological and phonetic assimilation (e.g., omitting

consonants in clusters like *st* or *zd* and sonorous/non-sonorous modifications). This approach ensured that the dictionary reflected natural pronunciation rather than just rigid orthography.

For Serbian ASR, Jakovljević and Pekar [141] proposed a unique tying procedure based on linguistic "levels of similarity" derived from the place and manner of articulation. They noted that previous (left) phonemes have more influence on leading states of a model, while successive (right) phonemes dominate ending states, allowing for robust estimation even with limited data.

Ipšić [142] introduced polyphone modeling for Slovenian, where a phone is modeled with an arbitrary length of context based on a minimal number of occurrences in the training set, allowing units to span over entire words if their frequency is high enough. To reduce reliance on expert linguistic knowledge, data-driven methods were introduced to generate broad phonetic classes based on confusion matrices [143]. Experiments on Slovenian showed that these data-driven classes often outperformed expert-defined ones, particularly in reducing WER for application words and isolated digits [132].

Further refinements included the use of sub-phonemes (separately modeling the closure and explosion of stops and affricates) and the conversion from phoneme-based to grapheme-based acoustic models [132]. Grapheme models are particularly advantageous for languages like Slovenian, where grapheme-to-phoneme rules are not yet fully standardized, as they eliminate the error introduced by inaccurate phonetic transcriptions.

5.2.4. Feature optimization and dimensionality reduction

High-dimensional feature spaces often suffer from poor generality in low-resource settings. Linear Discriminant Analysis (LDA) and Heteroscedastic LDA (HLDA) were employed to reduce dimensionality while maximizing class separability [144]. Studies on Serbian telephone data found that concatenating feature vectors across 7 successive frames and then projecting them to 32 or 35 dimensions using HLDA yielded the lowest word error rates. This configuration captured vital temporal information that standard delta and acceleration coefficients alone could not.

Vlaj et al. [131] investigated noise robustness, finding that frame attenuation (reducing spectral magnitude of noise frames) was more robust than frame dropping, as it was less sensitive to errors in Voice Activity Detection (VAD). They proposed a VAD based on log filter-bank magnitudes with a "hangover" criterion to prevent misclassifying weak fricatives as noise. In a major comparative study, Rotovnik et al. [145] compared HTK, ISIP, and Julius decoders on Slovenian LVCSR, finding that the Julius decoder offered a 65% improvement in the real-time factor while maintaining high accuracy, especially when using sub-word units. Research [146] comparing different Gaussian Mixture Model (GMM) structures (diagonal, semi-tied covariance, and full covariance) determined that for Serbian ASR, full covariance matrices achieved the best balance of accuracy and computational speed, despite their higher parameter count.

5.2.5. Cross-lingual transfer and resource-efficient bootstrapping

Given that many South Slavic languages are considered under-resourced, cross-lingual transfer became a vital strategy for rapid development, such as the COST 278 MASPER initiative [129]. The core idea was to bootstrap acoustic models using data from related languages to bypass the need for massive, manually transcribed target-language corpora [147].

The COST 278 MASPER initiative confirmed that language similarity is the dominant factor in cross-lingual performance; for example, Slovak models performed best for Slovenian due to their shared Slavic origin. Multilingual models trained on a pool of source languages (e.g., German, Spanish, and Slovak) were found to be more robust than monolingual source models when ported to an unseen target language [148]. Extreme cases explored building ASR without *any* target language text resources [149]. By using Croatian phoneme recognizers and Croatian phoneme-to-grapheme models, researchers were able to segment Slovene speech into word units and build an initial dictionary and language model directly from untranscribed audio and translations in other languages.

Žgank et al. [150] compared agglomerative (bottom-up) versus tree-based (top-down) triphone clustering for porting models to unseen languages, finding that tree-based systems typically performed better. For Croatian ASR, models were successfully trained "cost-free" by data mining broadcast news archives (HRT) [151]. By aligning loosely related summaries to audio using the Minimum Edit Distance (MED) algorithm, researchers generated substantial training data that yielded results comparable to models trained on expensive, dedicated databases.

5.2.6. Robustness and spontaneous speech phenomena

As ASR moved toward transcribing interviews and news, modeling spontaneous disfluencies became necessary. Žgank et al. [152] demonstrated that modeling filled pauses (e.g., "eee") and onomatopoeias using phonetic broad classes (grouping them as vowels or voiced consonants) improved Slovenian word accuracy by 5.70% without increasing recognition time. Finally, these classical HMM methods were applied in practical, distributed architectures. Kos et al. [153] presented the Genesis platform, which utilized word-based speaker-independent models in a distributed client/server architecture to enable natural human-machine interaction in intelligent Slovenian environments, achieving high accuracy with low command-to-operation delays.

As ASR systems moved from lab environments to real-world applications like Distributed Speech Recognition (DSR) and mobile services, environmental robustness became a priority. Researchers at the University of Maribor compared frame dropping (eliminating noise-only frames) and frame attenuation (reducing their spectral magnitude) [131]. Their findings consistently favored frame attenuation, as it was less sensitive to errors in Voice Activity Detection (VAD) and produced more stable acoustic modeling [154]. Beyond standard Fourier-based MFCCs, Wavelet Packet Parametrization was explored by Modić et al. [155] for its potential to provide better time localization of transient phenomena, showing improved robustness in variable noise conditions for both Slovenian and English tasks.

The transcription of spontaneous speech introduced the challenge of filled pauses (like "eee" or "uh") and onomatopoeias, which make up nearly 5% of tokens in Slovenian broadcast data. Modeling these interjections using phonetic broad classes (e.g., grouping them as vowels or voiced consonants) improved word accuracy by 5.7% without increasing recognition time [152].

5.2.7. Transition to neural architectures

While classical HMM-GMM systems were effective for read speech and limited domains, their reliance on diagonal covariance matrices and linear feature transformations struggled to capture the full variability of spontaneous, multi-dialectal Slavic speech [146], [156]. The persistence of the "dimensionality wall" and the inherent data sparsity of high-dimensional acoustic spaces for low-resource languages necessitated a shift in methodology [144].

As the field matured, the focus turned away from purely generative models toward discriminative training criteria and the integration of more powerful statistical estimators. This transition, which began with the use of Tandem systems and MLP-based bottleneck features, provided the bridge to the modern neural era. These advanced statistical refinements, which we will discuss in the next section, sought to replace the traditional GMM state-emission probabilities with the non-linear, discriminative power of deep learning, finally addressing the long-standing challenges of dialectal diversity and environmental uncertainty in South Slavic ASR [156].

5.3. Neural and end-to-end ASR

The transition from traditional statistical models to deep learning has fundamentally reshaped the landscape of ASR for South Slavic languages. Modern neural approaches have moved toward architectures that can directly model the non-linear complexities of speech while addressing the specific linguistic hurdles of Croatian, Serbian, Slovenian, and Bosnian. These hurdles, primarily high inflectivity, rich morphology, and the lack of massive, high-quality transcribed datasets, have necessitated the development of specialized neural strategies ranging from hybrid deep neural networks to pure end-to-end (E2E) transformer-based architectures.

5.3.1. From GMM-HMM Hybrids to deep neural acoustic models

The first phase of the neural revolution in the region focused on hybrid DNN-HMM systems, where deep neural networks (DNNs) replaced the GMM emission estimators while retaining HMM framework. During this time, acoustic modelling shifted from GMMs to DNNs within the HMM framework, while language modelling remained based on traditional statistical n-gram models.

Among the earliest demonstrations of this shift for the region was work on Serbian LVCSR using the Kaldi toolkit [157], which showed that DNN acoustic models, initialized with stacked restricted Boltzmann machines (RBMs) and trained using cross-entropy against triphone-state targets, consistently outperformed their GMM-HMM counterparts: a three-hidden-layer DNN achieved 1.86% WER on a 14,000-word test set, against 2.19% for the best

boosted-MMI GMM-HMM system, with further confirmation on a held-out set that used a zero-probability unigram language model to isolate pure acoustic model quality. A parallel practical deployment of the similar GMM-HMM paradigm was the Serbian Voice Assistant [158], an Android application that used noise-augmented MMI and boosted-MMI acoustic models trained in Kaldi; with noise at SNR 13 dB added to the training set, the boosted-MMI system reached 6.07% WER on clean speech, establishing early benchmarks for robust mobile deployment of South Slavic ASR.

The Kaldi-based DNN-HMM architecture was progressively refined through systematic ablation of network depth, neuron count, frame splicing strategy, and acoustic feature design. On a 215-hour Serbian corpus (audio books and mobile recordings), Pakoci et al. [159] showed that an eight-layer time-delay neural network (TDNN) with 625 neurons per layer and a mixed −1,0,1 / −3,0,3 splicing scheme emerged as the strongest configuration in terms of the WER–speed trade-off, yielding 9.45% WER; incorporating pitch features (weighted log-pitch, delta-log-pitch, and NCCF value) reduced this to 9.18%, and combining pitch with accent-specific vowel models for Serbian's five lexical accent categories pushed the best result to 9.06% WER and 2.37% CER. The gap between WER and CER — roughly a factor of four — was a consistent marker across Serbian experiments and pointed directly to the morphological fragility of the acoustic-model–n-gram pairing: the correct lemma was recognized but the wrong inflectional suffix was chosen, an issue that acoustic modeling alone could not solve.

Parallel work by Nouza et al. [160] covering the entire South Slavic family demonstrated that a single acoustic modeling pipeline could be extended across all seven languages (Croatian, Serbian, Bosnian, Montenegrin, Slovene, Macedonian, Bulgarian) using cross-lingual bootstrapping from Czech acoustic models, followed by lightly supervised iterative retraining on web-harvested broadcast speech. The DNN-HMM systems (five hidden layers with 1024–1024–768–768–512 neurons, trained via triphone GMMs using the Torch toolkit) achieved broadcast-news WER values between 16.8% and 21.5% (before excluding out-of-language passages), with Macedonian performing best at 14.54% after OOL exclusion. This work underscored both the potential of cross-lingual transfer across closely related languages and the ceiling imposed by the limited diversity of automatically harvested training data.

For Slovenian, the DNN-HMM trajectory followed a similar course but under more severe data constraints. A system [161] trained on 30 hours of BNSI broadcast news speech reached 31.39% WER with a DNN acoustic model (p-norm feed-forward, 2 hidden layers, 30 epochs), substantially below the 34.03% obtained with the GMM-HMM baseline. Crucially, DNN also showed greater robustness to domain mismatch, limiting WER degradation on the out-of-domain Interface emotional speech database to 35.27% versus 39.54% for HMMs. While emotional speech degraded the GMM-HMM baseline by 5–7 percentage points absolutely (specifically ranging from 6.14% to 7.02%, averaging 6.61%), the DNN acoustic model proved more robust, limiting absolute degradation to approximately 4.6% to 5.3% (averaging 4.91% absolute), with no significant difference observed between the emotions of disgust, joy, and sadness in terms of the magnitude of degradation.

5.3.2. Neural language modelling and morphological integration

While neural acoustic models brought steady improvement, the dominant source of remaining errors in South Slavic LVCSR resided in the language model. South Slavic languages are fusional and highly inflected languages, where a single lemma can realize dozens of distinct surface forms encoding case, number, gender, and tense, inflating effective vocabulary size and generating a large class of "weak substitution" errors — words that sound similar and share a lemma but differ in morphosyntactic function.

The transition toward neural language modeling in South Slavic languages began with recurrent neural network language models (RNNLMs). The classical n-gram approach with Kneser-Ney smoothing by Pakoci et al. [162] achieved WER of 6.90% and CER of only 2.20% with additional POS data considered. However, their RNNLM approach yielded WER of 4.34% and CER of 1.48%, demonstrating the advantage of RNNLMs over traditional n-gram language models. The RNNLM was implemented using the Kaldi-RNNLM framework, which supports subword letter n-gram features, word-frequency and word-length augmented features, and importance sampling to scale to large vocabularies.

A follow-up study [163] applied the same morphological-feature approach to an expanded RNNLM architecture and observed a similar improvement on a different test corpus, reducing WER from a 6.37% RNNLM baseline to 6.08% with the full set of morphological categories incorporated. The authors also compared four prior RNNLM implementations (Mikolov RNNLM, Faster-RNNLM, CUED-RNNLM, and TensorFlow LSTM) against a 3-gram baseline. All RNNLM variants showed substantial improvements over the 3-gram baseline, with the TensorFlow LSTM achieving 7.25% WER and 1.99% CER (approximately 20% relative reduction over 3-grams) on a 250,000-word vocabulary system. trained on a corpus of 1.4 million sentences (26 million words).

5.3.3. Transfer Learning and domain adaptation

Neural acoustic models trained on large general corpora can be adapted to narrow domains through transfer learning. This approach, first applied to Serbian ASR by Popović et al. [164], adapted an 8-layer TDNN trained on studio and audiobook speech to mobile voice assistant commands, achieving 4.44% WER, which is a 16.7% relative reduction from the 5.34% baseline, and with a corresponding 10.8% relative CER improvement. These improvements were partly attributed to the target domain's limited vocabulary (<4,000 words), which reduced the impact of morphological complexity.

A subsequent expanded study by Popović et al. [165] combined domain and environmental adaptation by training on a larger corpus that included both Serbian and Croatian audio (approximately 535 hours) and attempting to adapt the resulting acoustic model to noisy conditions by fine-tuning on pure noise recordings or a speech-plus-noise mixture. While adaptation using only noise recordings reduced word insertions (from around 10,000 to 4-5,000), it increased substitutions and deletions, resulting in higher WER that made it impractical as a standalone technique. In contrast, fine-tuning on speech-noise mixtures

improved noise robustness while largely preserving recognition accuracy, highlighting the trade-off between domain adaptation and generalization in under-resourced languages.

5.3.4. End-to-end architectures with CTC-based training

To overcome the complexity of modular hybrid systems, modern speech recognition shifted toward end-to-end (E2E) architectures. These unified networks eliminate the need for separate acoustic and pronunciation modules, mapping raw audio directly to text.

The same Serbian chain-training study [159] introduced earlier for its hybrid TDNN results also benchmarked two alternative architectures on the identical 215-hour database, allowing a direct comparison across paradigms. An early Eesen-based CTC system achieved 14.68% WER. While this was higher than the TDNN chain model (9.06%), it significantly outperformed the older GMM-HMM baseline (21.80%), proving that E2E models were already competitive even with limited training data. The same study also reported a hybrid LSTM-based acoustic model, using an LSTM with recurrent projection. It achieved the best overall accuracy at 9.00% WER, but required a massive computational cost, as training took nearly a full week (compared to 9.5 hours for the TDNN), and decoding speed dropped to 23% of real-time. Finally, introducing sequence-level chain training, which uses GPU acceleration and a simplified topology, allowed the hybrid system to easily surpass the end-to-end CTC results while keeping its standard decoding structure.

For Slovenian, Hailu et al. [166] tested a CNN-based end-to-end CTC system with audio-codec data augmentation technique, which includes re-encoding training audio from MP3 to a lower-bitrate Opus format to introduce spectral variation without harming intelligibility. On a 3-hour Slovenian Common Voice subset, this reduced CER from 21.56% to 20.35%, a gain consistent with the 1.05–2.78 point improvements the authors observed for Dutch, Turkish, and Amharic in the same study. They argue the technique is appealing for low-resource languages since it works on raw audio, requires no extra annotated data, and leaves feature extraction unchanged.

Not all deep learning approaches in this category are end-to-end in the strict sense – the final study reviewed here uses a modular, non-jointly-trained architecture, included to illustrate how deep learning has been applied to South Slavic ASR even outside the CTC/E2E paradigm. Fazlić et al. [167] addressed Bosnian audio-visual digit recognition, a task with no prior public dataset. They built a dedicated 150-sample corpus (expanded to 750 via noise augmentation) and proposed a dual-branch architecture: a CNN-RNN visual branch (GoogLeNet or ResNet-50 with LSTM) for lip reading, and a CNN audio branch on FFT spectrograms. Instead of being trained end-to-end, the two branches were trained separately and combined only during inference by summing their predictions. The audio-only model and the fused multimodal model both reached 100% accuracy, while the visual-only branch reached 72%. Despite the small scale, this demonstrates multimodal deep learning's applicability to under-resourced South Slavic languages, and points to jointly-trained end-to-end or transformer-based multimodal architectures as an unexplored direction for Bosnian.

5.3.5. Conformer and transformer end-to-end systems

The most decisive transition occurred with the transformer-based end-to-end architectures, particularly the conformer — a model combining self-attention with convolutional layers to capture both long-range dependencies and local acoustic patterns.

Žgank and Donaj [168] trained a conformer-based Slovenian recognizer with the WeNet toolkit on 597 hours of the Artur corpus, achieving 6.69% WER on the held-out Artur test set — a result the authors directly compare to English-level ASR benchmarks, and a major improvement over the 15.17% WER obtained with the prior hybrid GMM-DNN system. This confirms that sufficiently large corpora (≥500 hours) can overcome the morphological challenges of Slovenian in the end-to-end regime, at least for the speaking style represented in the training data. However, the same authors found this robustness was sharply domain-dependent. Testing the model on the out-of-distribution Interface emotional speech database, WER rose to 33.62% for neutral slow speech and 40.49% for neutral fast speech, while emotional styles averaged 38.08% (ranging 35.36% for surprise to 40.02% for anger). The authors traced this largely to speech-rate mismatch: the Artur test set averaged 8.09 characters/second versus 16.86 for the fast Interface subset, with almost no distributional overlap, compounded by the training corpus's complete lack of emotional or stylistic variation. This mirrors an earlier DNN-HMM finding that emotional speech raised WER by 5–7 points even for more flexible acoustic models [161].

Sepesy Maučec et al. [169] also applied transformer sequence-to-sequence models to Slovenian spoken language normalization: the conversion of literal, spoken-form transcriptions into standard written text. Both character-based LSTM and transformer models significantly outperformed statistical machine translation baselines, with the transformer reaching 3.00% WER and 0.90% CER on the aggregated test set versus 3.87% WER for the statistical baseline. LSTM and transformer results were comparable overall, though the transformer showed a measurable edge on the Gos and GosVL datasets while LSTM performed slightly better on the three Artur subsets. Error analysis revealed a qualitative advantage for the transformer: the errors it produced are easier to correct in post-editing, which is a practically important consideration for the costly process of building annotated speech corpora.

Having established that conformer and transformer architectures can reach strong in-domain performance, a natural next question is how well these same architecture families generalize to related but unseen language varieties. Zhang et al. [170] tested four Croatian ASR models on Čakavian, an endangered variety with distinct phonology, morphology, and vocabulary. A conformer-CTC model performed best (~39.4% WER), followed by wav2vec2-XLS-R with a language model (~45.2%), wav2vec2 alone (~45.6%), and Whisper-base (~55.9%) — differences driven mainly by training data volume rather than architecture. Whisper struggled most, likely due to limited Croatian data and a mismatched vocabulary. Errors clustered in two patterns: short function words absent from training vocabularies were often dropped, and word-endings accounted for 45–50% of substitutions, reflecting Čakavian's distinct inflection. The conformer's vocabulary also lacked key Čakavian tokens, breaking them into character bigrams, which is a gap the authors suggest targeted fine-tuning could fix.

The broader pattern across this body of work is consistent: for South Slavic languages, neural architectures at every level of the pipeline — acoustic modeling, language modeling, and full end-to-end systems — yield systematic improvements over their statistical predecessors, but the morphological richness of the language family and the scarcity of annotated data continue to define the ceiling. Hybrid DNN-HMM systems with morphologically informed RNNLMs represent the mature engineering baseline; conformer-based end-to-end systems have closed the gap with English-level performance under in-domain conditions; and self-supervised wav2vec2 fine-tuned models offer a pragmatic path for low-resource languages and varieties, provided that fine-tuning data aligns with the target acoustic and morphological distribution.

5.4. Challenges in South Slavic ASR

The studies discussed in Sections 5.2 and 5.3 often described complete ASR systems for South Slavic languages, treating obstacles such as inflection, data scarcity, or noise as background motivation for proposed architecture. The papers reviewed in this section instead mostly take these obstacles themselves as the primary object of study, quantifying how much accuracy is lost to channel mismatch or background noise, characterizing the scale of data sparsity in a real domain, or explicitly designing methods to counteract a named limitation. Read together, these studies offer a more direct picture of what makes South Slavic ASR difficult.

5.4.1. Linguistic complexity: morphology, prosody, language modelling

South Slavic languages are highly inflected, and this property creates difficulties across every part of an ASR system. A single noun stem can produce dozens of inflected surface forms across cases, numbers, and grammatical categories. Žgank [171] showed that this inflates vocabulary size and raises the out-of-vocabulary (OOV) rate in ASR systems: every missing word causes a substitution error, with language model errors adding a further factor of approximately 1.5x. Stem-ending language models reduce OOV rates but trade away n-gram prediction quality, since shorter units cover less linguistic context.

Morphological richness also affects acoustic modeling. Inflected forms from the same stem often differ by only a single final phoneme, making them acoustically confusable. Žgank tested the impact of this acoustic similarity by using a dedicated Slovenian highly inflected word database. Even under clean, noise-free conditions, speech recognition accuracy on highly inflected test sets fell from a non-inflected baseline of 93.3% to as low as 63.3% in the worst case, as the system frequently substituted highly similar endings. Under adverse conditions, such as background music, these inflected endings become heavily masked, causing further steep declines in word accuracy. A comparable pattern emerges in Serbian acoustic modeling [159]: time-domain neural network (TDNN) systems show a WER-to-CER gap of roughly a factor of four, indicating that the correct lemma is usually recognized but the wrong inflectional suffix is chosen, which is a failure mode that acoustic modeling alone cannot resolve.

Croatian adds prosodic complexity on top of inflectional complexity. Mikelić Preradović and Načinović Prskalo [172] noted that the language has four pitch-accent types,

that stress can fall on almost any syllable, and that it can shift across paradigms or onto preceding clitics. This makes a prosodic word on average 40% longer (3.12 syllables) than its corresponding orthographic word (2.25 syllables). Because standard digital dictionaries typically mark stress only on canonical lemmas, leaving inflected entries without phonetic detail, the authors proposed a hybrid approach. This system first applies morphological accentuation rules to inflected word forms and then utilizes machine learning (classification trees) to automatically assign stress to OOV words by identifying shared linguistic features. This hybrid approach achieves a final stress-assignment accuracy of 95.3% over a lexicon of 1,011,785 inflected word forms. Žgank [173] addressed the related grapheme-to-phoneme (G2P) problem differently: the UMB Broadcast News system bypasses phonemic modeling entirely and uses graphemes as acoustic units directly, avoiding G2P error propagation at the cost of flexibility for dialectal variation and non-standard orthography.

The scale of the lexicon problem is visible at the engineering level. Dobrišek et al. [174] found that a 200K-word Slovenian pronunciation dictionary required approximately 1.69 million finite-state transducer (FST) states before optimization, compared to 720K states for a 133K-word English dictionary. They noted that inflectional regularity does allow more state-sharing once minimized, but the initial lexicon size reflects how much more complex the South Slavic vocabulary space is to begin with.

Language modeling faces the same fundamental problem. Pakoci et al. [175] observed that the large number of morphological surface forms means many valid word sequences are unseen even in large training corpora, and standard n-gram smoothing cannot fully compensate. They showed that domain-specific settings make this worse — in Serbian medical text, practitioners switch between Serbian and Latin, and a single abbreviation such as *IV* can correspond to four distinct medical terms in either language, each with its own inflectional suffixes.

Ljubešić and Lauc [176] described BERTić a transformer-based language model pre-trained on 8.38 billion tokens of crawled web text across the closely related and mutually intelligible Bosnian, Croatian, Montenegrin, and Serbian domains. When evaluated on token classification tasks, BERTić achieved state-of-the-art performance on standard corpora and significantly outperformed prior multilingual and regional baselines (such as mBERT and cseBERT) on highly challenging non-standard social media text. Although BERTić was developed for text-based NLP tasks rather than ASR directly, it is relevant to speech recognition because transformer language models can be used to rescore ASR outputs or provide contextual language modeling in end-to-end speech recognition systems. This same morphosyntactic and dialectal fragility appears directly in speech recognition: an evaluation of Croatian ASR models on Čakavian [170] found that 45–50% of all substitution errors occurred at word endings, a direct consequence of inflectional divergence between the dialect and standard Croatian rather than a signal-processing failure.

5.4.2. Resource and acoustic constraints: data scarcity and acoustic variability

South Slavic languages are under-resourced by any standard comparison. Mikelić Preradović and Načinović Prskalo [172] described Croatian as lacking the corpora and resources needed for robust text and speech processing, noting that its first open ASR training dataset was published only recently (ParlaSpeech [103] in 2022). Žgank [171] found that Slovenian ASR systems underperform comparable Western European systems by approximately 15%, identifying limited training data as a key contributing factor. For Serbian, the publicly available speech databases are few, which makes data augmentation a practically important strategy: Galić and Grozdić [177] tested pitch shifting, time stretching, and volume control on isolated-word Serbian recognition, finding that pitch-shift augmentation yielded a statistically significant WER reduction for CNN-based models, while HMM and SVM systems showed no consistent benefit. Their result shows that the improvements from augmentation are architecture-dependent and cannot be assumed to generalize across model types.

Domain-specific data scarcity compounds the general resource problem. Pakoci et al. [175] worked with a 14.8 million-word corpus of Serbian medical findings, which is substantial by the standards of the language, that was still characterized by ambiguous abbreviations, informal language, and code-switching to Latin. They proposed class-based n-gram language modeling to manage this, achieving a perplexity of 59.55 compared to a range of 102–327 reported by other methods trained on up to 25 million words. Cross-lingual data pooling is a widely used mitigation, and Pakoci et al. showed that acoustic models trained on 904 hours of combined Serbian and Croatian speech outperformed Serbian-only models, and Ljubešić and Lauc [176] built BERTić on all four closely related Bosnian, Croatian, Montenegrin, and Serbian varieties precisely because none has sufficient standalone data on its own.

Signal-level degradation further limits robustness. Žgank [173] showed that the Slovenian broadcast news recognizer dropped from 97.12% accuracy on wide-band speech to 83.69% on narrow-band telephone speech — a 13.43 percentage-point absolute decrease. He noted that the F2 focus condition (narrow-band) was already identified as a known system weakness before this study, indicating that they were aware of the gap but lacked a quantified account of it. Žgank [171] also evaluated the impact of acoustic degradation on baseline speech recognition using clean studio recordings of isolated digits (93.88% baseline accuracy) and city names (98.53% baseline accuracy). When high-level instrumental music background was added, recognition accuracy dropped to as low as 23.58% for digits and 15.14% for city names. This decline is much larger than the roughly 20% performance decrease typically reported for comparable English broadcast news systems under degraded acoustic conditions. Beyond background noise, Žgank noted that spontaneous speech recognition in the region is further complicated by Slovenian's rich dialectal landscape, with more than 50 dialects, and frequent end-vowel reduction. These features introduce significant acoustic variability and make word forms more difficult to distinguish.

Speaking style is a related but distinct source of degradation: a DNN-HMM Slovenian system [161] showed WER rising by 5–7 percentage points absolutely under

emotional speech, and this gap widened sharply for a newer conformer-based end-to-end system [168], whose WER rose from 6.69% in-domain to an average of 38% under emotional and tempo-mismatched speech, suggesting that stronger architectures do not automatically confer robustness to speaking-style variation.

Spontaneous speech also introduces filled pauses (hesitations, prolonged sounds, and filler words) that are rarely transcribed and therefore absent from training data. In a study of filled pause detection across Slovenian, Croatian, and Serbian, Ljubešić et al. [178] found that detection models achieve F1 above 0.95 in-language but that false positives arise from language-specific phonetic features: nasals are the main confusion source in Slovenian and Croatian, while the conjunction *a* with a prolonged vowel causes the most errors in Serbian. They argued that these are not noise artifacts but structural features of natural speech that standard ASR pipelines for South Slavic languages do not yet account for.

6. SYNTHESIS OF RESULTS

The PRISMA screening process provides direct empirical confirmation of the gap motivating this review. Query 1 (compositional generalization and NLP/speech terms) and Query 2 (South Slavic identifiers and ASR/speech-resource terms) were designed around the expectation that these two literatures would show little to no overlap, and the screening results are consistent with that expectation: not a single record retrieved by Q1 also appeared in the Q2 result set, and none of the 76 CG/NLP studies included in the final corpus (Section 4) also appears among the 86 Croatian/South Slavic ASR studies (Section 5). Within the three databases searched (Web of Science, Scopus, IEEE Xplore) through May 2026, no study was found simultaneously indexed under both compositional generalization and South Slavic speech recognition terms, which is a search-level observation rather than definitive proof, since relevant work could exist under different terminology or outside these databases. Still, this absence lends additional support to the qualitative gap identified in Section 4.6 and motivates the future work.

Section 4 identifies three consistent findings. First, structure helps: models with tree- or graph-based architectural biases consistently outperform flat sequence models on compositional generalization tasks. Second, scale alone is not enough: both fine-tuning and in-context learning show little or even negative improvement in compositional generalization as model size increases, suggesting that this limitation has migrated to new models rather than being solved. Third, transfer is limited: improvements in multilingual and morphological generalization do not transfer automatically to agglutinative or low-resource languages. Importantly, these findings are only visible because the studies use dedicated diagnostic evaluation splits. For example, on COGS, the same model can achieve 96–99% accuracy on in-distribution examples but only 16–35% on compositional generalization examples, a difference that would remain hidden if only overall accuracy were reported. Table 6.1 provides an overview of the approaches, benchmarks and generalization findings reviewed in this section.

Section 5 reveals a similar pattern from the perspective of speech recognition. Again, structure helps: phonetic decision trees compensate for the fact that triphones cover only 12–22% of all theoretically possible combinations in Slavic languages. Scale alone is not sufficient: even large corpora such as ParlaSpeech-HR (1,816 hours) still require explicit morphological interventions like sub-word decomposition and lexical adaptation to control OOV growth. Transfer is possible but has not been evaluated compositionally: cross-lingual Slovenian-Croatian modeling and MASPER bootstrapping demonstrate real gains, but these gains are measured only using aggregate WER. Furthermore, every reviewed South Slavic ASR study reports performance exclusively via aggregate WER/CER on random or speaker-disjoint splits. No study separates examples containing previously seen morphological combinations from genuinely novel ones. As a result, an ASR system could simply memorize frequent word forms while performing poorly on unseen stem–suffix combinations, without this weakness appearing in the reported results.

Table 6.1. Overview of approaches, benchmarks, and generalization findings in compositional generalization research.

Approach	Core description	Main benchmarks	Representative studies	Performance & Generalization Findings
Synthetic CG benchmarks	Engineered distribution gaps between training and test sets to isolate compositional abilities from random memorization.	SCAN, COGS, CFQ, gSCAN, SyGNS, CTL++	SCAN (2018) [1]; COGS (2020) [2]; Keysers et al. (2019) [8]	Standard sequence models score high in-distribution (96–99% on COGS; >95% on CFQ random splits) but drop catastrophically on OOD splits (16–35% on COGS; <20% on CFQ maximum-divergence splits).
Architecture with structural bias	Modifying neural network architectures or introducing inductive biases (trees, graphs, group equivariance) to better match linguistic structures.	SCAN, COGS, CFQ	Permutation Equivariant Networks (2019) [42] ; Syntax-Semantics Factorization (2020) [52]; Hierarchical Poset Decoding (2020) [49]; Grounded Graph Decoding (2021) [50]	Equivariant seq2seq architectures significantly outperform regular seq2seq and convolutional baselines on SCAN. Factorizing alignment and translation components allows known grammatical rules to apply systematically to novel words. Hierarchical poset decoding enforces partial permutation invariance by representing language understanding as a DAG-structured partially ordered set . Grounded Graph Decoding solves CFQ's MCD1 split with 98% accuracy, improving generalization over flat-embedding baselines.
Training / data augmentation	Altering data and training procedures via automated rule-based recombination, grammar sampling, or consistency regularization, without modifying the underlying model architecture.	COGS, SCAN, GeoQuery, SMCalFlow-CS, CoGnition	Compositional Structure Learner (CSL, 2022) [61]; TreeMix (2022) [62]; LEXSYM (2023) [69]; LLM-based data augmentation (2023) [63]	Injects compositional priors into otherwise unstructured models and substantially narrows generalization gaps. CSL achieves 99.5% generalization accuracy on COGS and improves performance on SCAN, GeoQuery, and SMCalFlow-CS. LexSym reduces error on CLEVR-COGENT by roughly 33% and matches or surpasses task-specific baselines on COGS and SCAN.
Structured semantic tasks	Real-world semantic parsing, text-to-SQL, and KBQA tasks where natural language maps to executable database or knowledge graph structures.	GeoQuery, Spider-CG, CoSQL, SParC, LC-QuAD.	Spider-CG (2022) [76]; CERG-KBQA (2024) [79]; QDTQA (2023) [80]	Models suffer major performance drops on compositional splits (e.g., Spider-CG shows degradation even on clause-level sub-sentences seen during training); decomposing questions into substructures or using constraint-aware decoding helps partially recover generalization, with QDTQA achieving an 11% gain and new state of the art on LC-QuAD 1.0.

LLM-based approaches	Evaluating parameter scaling, in-context learning (ICL), prompt optimization (CoT, PoT), and fine-tuning in large pre-trained language models.	CFQ, SCAN, GeoQuery, MATH, GSM8K, Ordered CommonGen, CompoST	Hosseini et al. (2022) [81]; Qiu et al. (2022) [60] ; CompoST (2026) [88]	Scale and in-context learning narrow but don't close the CG gap, and LLMs remain well below human performance on harder tasks, confirming that scale and prompting alone haven't solved systematic compositionality.
Morphology-focused CG	Investigating morphological productivity and systematicity by treating morphemes as compositional primitives (inflection and derivation).	CsNoFEVER (Czech), Europarl (DBCA), Turkish/Finnish morphology tasks	DBCA (2023) [92]; Ismayilzada et al. (2025) [22]; Vrabcová and Sojka (2024) [91]	State-of-the-art multilingual LLMs (e.g., GPT-4, Gemini) struggle with morphological compositional generalization on novel word roots; Czech models show drastic accuracy drops on negated forms due to fusional morphology; smaller subword vocabularies improve generalization to unseen inflections in machine translation.
Multilingual / spoken domain	Evaluating cross-lingual transfer, structural differences across languages, and spoken language understanding (SLU) pipelines.	MCWQ, ATIS, SNIPS	Multilingual Compositional Generalization with Translated Datasets (2023) [94]; CG in SLU (2023) [21]	Multilingual models experience cross-lingual structural bottlenecks, and SLU models trained on clean transcripts have not been tested for acoustic or morphological robustness.

One particularly important similarity is that both fields face an evaluation problem. Research on compositional generalization introduced diagnostic evaluation splits because overall accuracy cannot distinguish memorization from true generalization. South Slavic ASR has not, to our knowledge, developed an equivalent evaluation methodology and therefore cannot currently measure this distinction. Text-based compositional generalization studies explicitly report the large drop in performance between in-distribution and out-of-distribution examples, such as 96–99% versus 16–35% on COGS and 99.98% versus 92.58% on the most difficult CFQ splits. In contrast, South Slavic ASR studies report substantial differences in WER and CER, including the fourfold variation discussed in Section 5.3.1, but these results provide only indirect evidence of a similar phenomenon because none of the reviewed studies evaluates systems on test sets organized by morphological novelty.

In combination, these findings reveal a similar pattern across both research areas: structural approaches consistently improve generalization, increasing model size alone does not solve the problem, and transfer across languages is possible but limited. This suggests that morphology-aware compositional evaluation, which has received little attention in South Slavic ASR, could be a promising next step. However, such evaluation requires a speech domain in which compositional structure is both linguistically well-defined and easy to evaluate. Numeral expressions are one suitable candidate because they are productively formed from a small set of primitives (units, teens, tens, hundreds, etc.) under clear morphological rules. This makes it possible to create evaluation splits that test novel but grammatically valid

combinations, rather than relying only on speaker- or utterance-based splits. Moreover, because numeral expressions also vary naturally in length, from single-digit numbers to complex multi-digit combinations, they are equally well suited to testing productivity, i.e., generalization to longer numeral sequences than those seen during training, alongside systematic recombination of familiar components. A dataset designed in this way could help reveal generalization weaknesses in South Slavic ASR that standard evaluation methods may be missing, while also providing a testbed for adapting text-based compositional generalization methods to the morphological complexity of South Slavic languages.

Table 6.2. Overview of ASR architectures, design features, and performance results across South Slavic languages.

Architecture	Key design architecture	Representative studies	Languages	Performance Results
Classical HMM-GMM	Context-dependent triphone modeling with decision-tree state tying. Incorporates explicit morphological decomposition (stems/endings) and class LMs to control OOV explosion.	Petek (1995) [127]; Imperl (1999) [128]; Croatian Large Vocabulary ASR (2011) [139]; LVCSR (2007) [18]	Croatian, Serbian, Slovenian	Petek's Slovenian digit recognition: 97.3% accuracy with gender-specific HMMs and MFCC features. Subsequent LVCSR work shows that only a small fraction of theoretically possible triphone combinations are observed in training, motivating decision-tree tying, and that explicit morphological decomposition in Slovenian LVCSR reduces OOV rates from around 6% to below 1%
Hybrid DNN-HMM & Neural Language Modeling	DNN emission estimators replacing GMMs. Extended with TDNN sequence-chain training, LF-MMI, interpolated neural/class-based LMs, and cross-domain transfer learning.	Kaldi DNN transfer learning (2019) [164]; Sequence-trained TDNN (2018) [159]; cross-lingual bootstrapping (2016) [160]	Serbian, Slovenian, 7 South Slavic languages	Serbian TDNN chain model achieves 9.06% WER on a multi-domain speech corpus, versus 21.80% for a GMM–HMM baseline. Interpolated RNNLM + 3-gram LM further reduces WER to 6.37%. Domain and environment adaptation via transfer learning yields about 16.7% relative WER reduction on mobile smartphone commands, whereas environmental noise adaptation on its own actually increases overall error rates while successfully limiting background insertion errors.
CTC-Based End-to-End & Multimodal Networks	Direct acoustic-to-text mapping without explicit HMM alignments, enabling audio-visual and multimodal fusion capabilities.	Eesen CTC (2018) [159]; CNN-CTC audio-codec augmentation (2020)[166]; Bosnian digits multimodal network (2024) [167]	Serbian, Slovenian, Bosnian	Eesen CTC achieves around 14.68% WER on Serbian LVCSR experiments, significantly outperforming the traditional HMM-GMM baseline of 21.80%. Bosnian audio-visual CNN–RNN models reach 100% accuracy on digit recognition in a constrained audio-visual setup, illustrating multimodal benefits on small vocabulary tasks.

Conformer & Speech Transformers	Self-attention combined with convolution for acoustic patterns (Conformer-CTC), massive self-supervised pretraining (wav2vec 2.0), and sub-word tokenization.	Slovenian E2E ASR (2025) [179]; transformer normalization (2024) [169]; Čakavian evaluation (2024) [170]	Slovenian, Croatian (incl. Čakavian)	Slovenian Conformer-CTC trained on ARTUR achieves 6.69% WER in-domain but degrades to 33.62% WER on out-of-domain neutral slow speech, and averages 38.08% WER on out-of-domain emotional. For Čakavian dialect transcription, Croatian ASR models face major out-of-domain generalization hurdles: while wav2vec 2.0 (with a 5-gram LM) and Conformer-CTC reach strong in-domain baselines of 4.3% and 4.7% WER on standard Croatian, their error rates drop to 45.2% and 39.4% mean WER respectively when evaluated on the Čakavian test set

7. CONCLUSION

Compositional generalization — a model's ability to recombine known elements in unseen ways while preserving meaning — has been studied intensively in recent years, with benchmarks such as SCAN, COGS, and CFQ motivating a wide range of architectural and training interventions. Yet the evidence assembled in this review shows the underlying systematicity gap has persisted rather than closed, resurfacing in each new generation of models from sequence-to-sequence architectures to large language models. This body of work also remains concentrated on English text, leaving open whether the same failure mode holds for morphologically rich languages and for speech, where it has barely been tested at all.

Automatic speech recognition for Croatian and related South Slavic languages has, by contrast, made substantial progress in recent years, from classical HMM-GMM systems adapted for high inflection to modern neural and end-to-end architectures trained on growing speech corpora. This progress has repeatedly confronted morphological richness as an engineering problem, through sub-word decomposition, lexical adaptation, and cross-lingual transfer. However, none of this progress has been evaluated from a compositional generalization perspective: performance is reported almost exclusively through aggregate WER and CER, metrics that cannot distinguish a system that has learned productive morphological structure from one that has simply memorized frequent word forms.

These two trajectories directly motivate the two research questions guiding this review. The first research question (RQ1) asked what approaches and benchmarks exist for compositional generalization in text and speech, and what limits them. The review found that compositional generalization has been extensively benchmarked (SCAN, COGS, CFQ) and addressed through architectural and training interventions, but that this body of work is concentrated on English text, with only a handful of studies extending to morphologically rich languages and only a single study extending to speech. None of the reviewed CG work addresses morphologically rich speech directly, and existing SLU-focused CG work does not engage with inflection, agglutination, or acoustic-morphological interaction.

The second research question (RQ2) asked what speech resources, models, and evaluation practices exist for Croatian and related low-resource Slavic languages. Here the review identified a substantial but uneven body of Croatian, Serbian, Slovenian, and Bosnian ASR research, spanning speech corpora (from GlobalPhone and SpeechDat(II) to ParlaSpeech-HR), classical HMM-GMM systems adapted for high inflection through sub-word modeling and lexical adaptation, and emerging neural/end-to-end approaches. This literature demonstrates that morphological richness is a dominant, well-documented engineering challenge for South Slavic ASR, but evaluation has remained centered on aggregate WER/CER rather than on whether systems generalize compositionally to unseen morpheme combinations.

Considered together, these findings point to a research gap at the intersection of compositional generalization, speech recognition, and morphological richness. Further research will explore this gap by adapting compositional generalization evaluation methods and morphology-aware modeling approaches developed for text to Croatian and other South Slavic ASR, with the aim of addressing it directly. As a starting point, numeral expressions

provide a suitable case for constructing compositional evaluation splits because they combine productively from a small, well-defined set of morphological primitives. They can therefore serve as an initial testbed before extending the approach to other compositional phenomena.

This review has several limitations. The literature search was restricted to three databases (Web of Science, Scopus, IEEE Xplore) and to full text publications available in English or Croatian. As a result, relevant studies published in other South Slavic languages or in venues not covered by these databases might have been missed. The empirical ASR scope was also deliberately restricted to Croatian, Serbian, Slovenian, and Bosnian, excluding Bulgarian and Macedonian due to their more limited resources and typological differences. Finally, because research at the intersection of compositional generalization and morphologically rich speech is still at an early stage, this review necessarily brings together findings from research areas that have not often been considered together. Some of the connections identified in this way are theoretically motivated but remain to be empirically validated by the dissertation research that follows.

LITERATURE

[1] B. Lake and M. Baroni, “Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks,” in *Proceedings of the 35th International Conference on Machine Learning*, PMLR, Jul. 2018, pp. 2873–2882. [Online]. Available: https://proceedings.mlr.press/v80/lake18a.html

[2] N. Kim and T. Linzen, “COGS: A Compositional Generalization Challenge Based on Semantic Interpretation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Nov. 2020, pp. 9087–9105. doi: 10.18653/v1/2020.emnlp-main.731.

[3] M. J. Page *et al.*, “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews,” *BMJ*, vol. 372, 2021.

[4] R. Montague, “Universal grammar,” *Theoria*, vol. 36, no. 3, pp. 373–398, 1970, doi: 10.1111/j.1755-2567.1970.tb00434.x.

[5] T. M. V. Janssen, “Compositionality: Its Historic Context,” in *The Oxford Handbook of Compositionality*, M. Werning, W. Hinzen, and E. Machery, Eds., Oxford: Oxford University Press, 2012, pp. 19–46.

[6] Z. An and W. Du, “Representational Homomorphism Predicts and Improves Compositional Generalization In Transformer Language Model,” Mar. 24, 2026, *arXiv*: arXiv:2601.18858. doi: 10.48550/arXiv.2601.18858.

[7] J. A. Fodor and Z. W. Pylyshyn, “Connectionism and cognitive architecture: A critical analysis,” *Cognition*, vol. 28, no. 1–2, pp. 3–71, 1988.

[8] D. Keysers *et al.*, “Measuring Compositional Generalization: A Comprehensive Method on Realistic Data,” in *Proceedings of the International Conference on Learning Representations*, Sep. 2019. [Online]. Available: https://openreview.net/forum?id=SygcCnNKwr

[9] K. Sun, A. Williams, and D. Hupkes, “The Validity of Evaluation Results: Assessing Concurrence Across Compositionality Benchmarks,” in *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, Singapore: Association for Computational Linguistics, Dec. 2023, pp. 274–293. doi: 10.18653/v1/2023.conll-1.19.

[10] D. Hupkes, V. Dankers, M. Mul, and E. Bruni, “Compositionality decomposed: how do neural networks generalise?,” Feb. 25, 2020, *arXiv*: arXiv:1908.08351. doi: 10.48550/arXiv.1908.08351.

[11] K. McCurdy, P. Soulos, P. Smolensky, R. Fernandez, and J. Gao, “Toward Compositional Behavior in Neural Models: A Survey of Current Views,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 9323–9339. doi: 10.18653/v1/2024.emnlp-main.524.

[12] M. Baroni, “Linguistic generalization and compositionality in modern artificial neural networks,” *Phil. Trans. R. Soc. B*, vol. 375, no. 1791, p. 20190307, Dec. 2019, doi: 10.1098/rstb.2019.0307.

[13] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989, doi: 10.1109/5.18626.

[14] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980, doi: 10.1109/TASSP.1980.1163420.

[15] G. Hinton *et al.*, “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012, doi: 10.1109/MSP.2012.2205597.

[16] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, in ICML ’06. New York, NY, USA: Association for Computing Machinery, Jun. 2006, pp. 369–376. doi: 10.1145/1143844.1143891.

[17] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” in *Soviet Physics Doklady*, Soviet Union, 1966, pp. 707–710.

[18] T. Rotovnik, M. S. Maučec, and Z. Kačič, “Large vocabulary continuous speech recognition of an inflected language using stems and endings,” *Speech Communication*, vol. 49, no. 6, pp. 437–452, Jun. 2007, doi: 10.1016/j.specom.2007.02.010.

[19] V. A. Friedman, “Macedonian,” in *The Slavonic Languages*, B. Comrie and G. G. Corbett, Eds., London: Routledge, 1993, pp. 249–305.

[20] E. A. Scatton, “Bulgarian,” in *The Slavonic Languages*, B. Comrie and G. G. Corbett, Eds., London: Routledge, 1993, pp. 188–248.

[21] A. Ray, Y. Shen, and H. Jin, “Compositional Generalization in Spoken Language Understanding,” in *Proceedings of Interspeech 2023*, 2023, pp. 750–754. doi: 10.21437/Interspeech.2023-1419.

[22] M. Ismayilzada *et al.*, “Evaluating Morphological Compositional Generalization in Large Language Models,” in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 1270–1305. doi: 10.18653/v1/2025.naacl-long.59.

[23] M. Ouzzani, H. Hammady, Z. Fedorowicz, and A. Elmagarmid, “Rayyan—a web and mobile app for systematic reviews,” *Systematic Reviews*, vol. 5, no. 1, p. 210, 2016, doi: 10.1186/s13643-016-0384-4.

[24] A. Petit, C. Corro, and F. Yvon, “Structural generalization in COGS: Supertagging is (almost) all you need,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1089–1101. doi: 10.18653/v1/2023.emnlp-main.69.

[25] A. Jabbar, C. Condoravdi, and C. Potts, “Distinguishing fair from unfair compositional generalization tasks,” in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 20796–20807. doi: 10.18653/v1/2025.findings-emnlp.1133.

[26] D. Tsarkov, T. Tihon, N. Scales, N. Momchev, D. Sinopalnikov, and N. Schärli, “*-CFQ: Analyzing the Scalability of Machine Learning on a Compositional Task,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, pp. 9949–9957, May 2021, doi: 10.1609/aaai.v35i11.17195.

[27] T. Gao, Q. Huang, and R. Mooney, “Systematic Generalization on gSCAN with Language Conditioned Embedding,” in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 491–503. doi: 10.18653/v1/2020.aacl-main.49.

[28] H. Yanaka, K. Mineshima, and K. Inui, “SyGNS: A Systematic Generalization Testbed Based on Natural Language Semantics,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, Aug. 2021, pp. 103–119. doi: 10.18653/v1/2021.findings-acl.10.

[29] R. Csordás, K. Irie, and J. Schmidhuber, “CTL++: Evaluating Generalization on Never-Seen Compositional Patterns of Known Functions, and Compatibility of Neural Representations,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 9758–9767. doi: 10.18653/v1/2022.emnlp-main.662.

[30] A. Liška, G. Kruszewski, and M. Baroni, “Memorize or generalize? Searching for a compositional RNN in a haystack,” Jul. 25, 2018, *arXiv*: arXiv:1802.06467. doi: 10.48550/arXiv.1802.06467.

[31] C. Wang, R. van Noord, A. Bisazza, and J. Bos, “Evaluating Text Generation from Discourse Representation Structures,” in *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, Association for Computational Linguistics, Aug. 2021, pp. 73–83. doi: 10.18653/v1/2021.gem-1.8.

[32] E. Goodwin, K. Sinha, and T. J. O’Donnell, “Probing Linguistic Systematicity,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Jul. 2020, pp. 1958–1969. doi: 10.18653/v1/2020.acl-main.177.

[33] E. Manino, J. Rozanova, D. Carvalho, A. Freitas, and L. Cordeiro, “Systematicity, Compositionality and Transitivity of Deep NLP Models: a Metamorphic Testing Perspective,” in *Findings of the Association for Computational Linguistics: ACL 2022*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2355–2366. doi: 10.18653/v1/2022.findings-acl.185.

[34] J. Valvoda, N. Saphra, J. Rawski, A. Williams, and R. Cotterell, “Benchmarking Compositionality with Formal Languages,” in *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 6007–6018. [Online]. Available: https://aclanthology.org/2022.coling-1.525/

[35] S. Wold, L. G. G. Charpentier, and É. Simon, “Systematic Generalization in Language Models Scales with Information Entropy,” in *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 1807–1819. doi: 10.18653/v1/2025.findings-acl.90.

[36] D. Sachan, M. Zaheer, and R. Salakhutdinov, “Investigating the Working of Text Classifiers,” in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 2120–2131. [Online]. Available: https://aclanthology.org/C18-1180/

[37] I. Dasgupta, D. Guo, S. J. Gershman, and N. D. Goodman, “Analyzing Machine-Learned Representations: A Natural Language Case Study,” *Cognitive Science*, vol. 44, no. 12, p. e12925, 2020, doi: 10.1111/cogs.12925.

[38] Z. Guo *et al.*, “Quantifying Compositionality of Classic and State-of-the-Art Embeddings,” in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 22130–22146. doi: 10.18653/v1/2025.findings-emnlp.1206.

[39] X. Fu and A. Frank, “The Mystery of Compositional Generalization in Graph-based Generative Commonsense Reasoning,” in *Findings of the Association for Computational*

Linguistics: EMNLP 2024, Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 8376–8394. doi: 10.18653/v1/2024.findings-emnlp.492.

[40] Y. Yao and A. Koller, "Structural generalization is hard for sequence-to-sequence models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5048–5062. doi: 10.18653/v1/2022.emnlp-main.337.

[41] L. Galke, Y. Ram, and L. Raviv, "Deep neural networks and humans both benefit from compositional language structure," *Nat Commun*, vol. 15, no. 1, p. 10816, Dec. 2024, doi: 10.1038/s41467-024-55158-1.

[42] J. Gordon, D. Lopez-Paz, M. Baroni, and D. Bouchacourt, "Permutation equivariant models for compositional generalization in language," in *International Conference on Learning Representations*, 2019.

[43] S. Basu *et al.*, "Equi-Tuning: Group Equivariant Fine-Tuning of Pretrained Models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, pp. 6788–6796, Jun. 2023, doi: 10.1609/aaai.v37i6.25832.

[44] S. Basu *et al.*, "Efficient equivariant transfer learning from pretrained models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 4213–4224, 2023.

[45] S. Ontañón, J. Ainslie, Z. Fisher, and V. Cvicek, "Making Transformers Solve Compositional Tasks," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 3591–3607. doi: 10.18653/v1/2022.acl-long.251.

[46] N. Patel and J. Flanigan, "Forming Trees with Treeformers," in *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, Sep. 2023, pp. 836–845. Accessed: Jul. 24, 2026. [Online]. Available: https://aclanthology.org/2023.ranlp-1.90/

[47] B. Wang, M. Lapata, and I. Titov, "Structured Reordering for Modeling Latent Alignments in Sequence Transduction," in *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, in Advances in Neural Information Processing Systems. La Jolla: Neural Information Processing Systems (NIPS), 2021.

[48] M. Lindemann, A. Koller, and I. Titov, "SIP: Injecting a Structural Inductive Bias into a Seq2Seq Model by Simulation," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 6570–6587. doi: 10.18653/v1/2024.acl-long.355.

[49] Y. Guo, Z. Lin, J.-G. Lou, and D. Zhang, "Hierarchical poset decoding for compositional generalization in language," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6913–6924, 2020.

[50] Y. Gai, P. Jain, W. Zhang, J. Gonzalez, D. Song, and I. Stoica, "Grounded Graph Decoding improves Compositional Generalization in Question Answering," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1829–1838. doi: 10.18653/v1/2021.findings-emnlp.157.

[51] L. Bergen, T. O'Donnell, and D. Bahdanau, "Systematic generalization with edge transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1390–1402, 2021.

[52] J. Russin, J. Jo, R. O'Reilly, and Y. Bengio, "Compositional Generalization by Factorizing Alignment and Translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Association for Computational Linguistics, Jul. 2020, pp. 313–327. doi: 10.18653/v1/2020.acl-srw.42.

[53] A. K. Chakravarthy, J. L. Russin, and R. O'Reilly, "Systematicity Emerges in Transformers when Abstract Grammatical Roles Guide Attention," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, Jul. 2022, pp. 1–8. doi: 10.18653/v1/2022.naacl-srw.1.

[54] R. Riveland and A. Pouget, "Natural language instructions induce compositional generalization in networks of neurons," *Nat Neurosci*, vol. 27, no. 5, pp. 988–999, May 2024, doi: 10.1038/s41593-024-01607-5.

[55] P. C. R. Lane and J. B. Henderson, "Towards Effective Parsing with Neural Networks: Inherent Generalisations and Bounded Resource Effects," *Applied Intelligence*, vol. 19, no. 1, pp. 83–99, Jul. 2003, doi: 10.1023/A:1023820807862.

[56] S. L. Frank, "Learn more by training less: systematicity in sentence processing by recurrent networks," *Connection Science*, vol. 18, no. 3, pp. 287–302, 2006.

[57] I. Farkaš and M. W. Crocker, "Syntactic systematicity in sentence processing with a recurrent self-organizing network," *Neurocomputing*, vol. 71, no. 7–9, pp. 1172–1179, 2008.

[58] Z. Zhang, P. Lin, Z. Wang, Y. Zhang, and Z.-Q. John Xu, "Complexity Control Facilitates Reasoning-Based Compositional Generalization in Transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 4, pp. 4336–4349, Apr. 2026, doi: 10.1109/TPAMI.2025.3646483.

[59] Y. Li, L. Zhao, J. Wang, and J. Hestness, "Compositional Generalization for Primitive Substitutions," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4293–4302. doi: 10.18653/v1/D19-1438.

[60] L. Qiu *et al.*, "Evaluating the Impact of Model Scale for Compositional Generalization in Semantic Parsing," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 9157–9179. doi: 10.18653/v1/2022.emnlp-main.624.

[61] L. Qiu *et al.*, "Improving Compositional Generalization with Latent Structure and Data Augmentation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 4341–4362. doi: 10.18653/v1/2022.naacl-main.323.

[62] L. Zhang, Z. Yang, and D. Yang, "TreeMix: Compositional Constituency-based Data Augmentation for Natural Language Understanding," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 5243–5258. doi: 10.18653/v1/2022.naacl-main.385.

[63] Y. Fang, X. Li, S. Thomas, and X. Zhu, “ChatGPT as Data Augmentation for Compositional Generalization: A Case Study in Open Intent Detection,” in *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, 2023, pp. 13–33. [Online]. Available: https://aclanthology.org/2023.finnlp-1.2/

[64] K. Wei, S. Ghosh, and S. Srivastava, “Compositional Generalization for Kinship Prediction through Data Augmentation,” in *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 13–19. doi: 10.18653/v1/2022.wnu-1.2.

[65] K. Owoeye, “Curriculum Compositional Continual Learning for Neural Machine Translation,” *Procedia Computer Science*, vol. 222, pp. 167–176, Jan. 2023, doi: 10.1016/j.procs.2023.08.154.

[66] X. Fu and A. Frank, “Exploring Continual Learning of Compositional Generalization in NLI,” *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 912–932, Aug. 2024, doi: 10.1162/tacl_a_00680.

[67] C. Liu *et al.*, “Learning Algebraic Recombination for Compositional Generalization,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Association for Computational Linguistics, Aug. 2021, pp. 1129–1144. doi: 10.18653/v1/2021.findings-acl.97.

[68] E. Akyurek and J. Andreas, “Lexicon Learning for Few Shot Sequence Modeling,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, Aug. 2021, pp. 4934–4946. doi: 10.18653/v1/2021.acl-long.382.

[69] E. Akyurek and J. Andreas, “LexSym: Compositionality as Lexical Symmetry,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 639–657. doi: 10.18653/v1/2023.acl-long.38.

[70] S. Kim, J. Kim, and K. Jung, “Compositional Generalization via Parsing Tree Annotation,” *IEEE Access*, vol. 9, pp. 24326–24333, 2021, doi: 10.1109/ACCESS.2021.3055513.

[71] F. Shen, Y. Yin, Y. Li, D. F. Wong, and Y. Zhang, “Consistency Regularization for Translational Compositional Generalization,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 4361–4374, 2025, doi: 10.1109/TASLPRO.2025.3622897.

[72] R. Chaabouni, R. Dessì, and E. Kharitonov, “Can Transformers Jump Around Right in Natural Language? Assessing Performance Transfer from SCAN,” in *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 136–148. doi: 10.18653/v1/2021.blackboxnlp-1.9.

[73] M. Damonte and E. Monti, “One Semantic Parser to Parse Them All: Sequence to Sequence Multi-Task Learning on Semantic Parsing Datasets,” in *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, Association for Computational Linguistics, Aug. 2021, pp. 173–184. doi: 10.18653/v1/2021.starsem-1.16.

[74] P. Shaw, M.-W. Chang, P. Pasupat, and K. Toutanova, “Compositional Generalization and Natural Language Variation: Can a Semantic Parsing Approach Handle Both?,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1:*

Long Papers), Association for Computational Linguistics, Aug. 2021, pp. 922–938. doi: 10.18653/v1/2021.acl-long.75.

[75] V. Dankers, E. Bruni, and D. Hupkes, "The Paradox of the Compositionality of Natural Language: A Neural Machine Translation Case Study," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4154–4175. doi: 10.18653/v1/2022.acl-long.286.

[76] Y. Gan, X. Chen, Q. Huang, and M. Purver, "Measuring and Improving Compositional Generalization in Text-to-SQL via Component Alignment," in *Findings of the Association for Computational Linguistics: NAACL 2022*, Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 831–843. doi: 10.18653/v1/2022.findings-naacl.62.

[77] A. Arora, S. Bhaisaheb, H. Nigam, M. Patwardhan, L. Vig, and G. Shroff, "Adapt and Decompose: Efficient Generalization of Text-to-SQL via Domain Adapted Least-To-Most Prompting," in *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, Singapore: Association for Computational Linguistics, Dec. 2023, pp. 25–47. doi: 10.18653/v1/2023.genbench-1.3.

[78] S. Hari Krishnan Parthasarathi, L. Zeng, and D. Hakkani-Tür, "Conversational Text-to-SQL: An Odyssey into State-of-the-Art and Challenges Ahead," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096170.

[79] J. Chen, H. Gao, Y. Wang, and X. Wu, "CERG-KBQA: Advancing KBQA-Systems through Classification, Enumeration, Ranking, and Generation Strategies," in *2024 IEEE Smart World Congress (SWC)*, Dec. 2024, pp. 1352–1359. doi: 10.1109/SWC62898.2024.00210.

[80] X. Huang, S. Cheng, Y. Shu, Y. Bao, and Y. Qu, "Question Decomposition Tree for Answering Complex Questions over Knowledge Bases," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 12924–12932, Jun. 2023, doi: 10.1609/aaai.v37i11.26519.

[81] A. Hosseini, A. Vani, D. Bahdanau, A. Sordoni, and A. Courville, "On the Compositional Generalization Gap of In-Context Learning," in *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 272–280. doi: 10.18653/v1/2022.blackboxnlp-1.22.

[82] S. Han and S. Padó, "Towards Understanding the Relationship between In-context Learning and Compositional Generalization," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia: ELRA and ICCL, May 2024, pp. 16664–16679. [Online]. Available: https://aclanthology.org/2024.lrec-main.1449/

[83] J. Zhao, J. Tong, Y. Mou, M. Zhang, Q. Zhang, and X. Huang, "Exploring the Compositional Deficiency of Large Language Models in Mathematical Reasoning Through Trap Problems," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 16361–16376. doi: 10.18653/v1/2024.emnlp-main.915.

[84] M. Zhang, J. He, S. Lei, M. Yue, L. Wang, and C.-T. Lu, "Can LLM Find the Green Circle? Investigation and Human-Guided Tool Manipulation for Compositional Generalization," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal*

Processing (ICASSP), Apr. 2024, pp. 11996–12000. doi: 10.1109/ICASSP48485.2024.10446355.

[85] F. Petruzzellis, A. Testolin, and A. Sperduti, "Benchmarking GPT-4 on Algorithmic Problems: A Systematic Evaluation of Prompting Strategies," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia: ELRA and ICCL, May 2024, pp. 2161–2177. [Online]. Available: https://aclanthology.org/2024.lrec-main.195/

[86] S. Lee *et al.*, "Reasoning abilities of large language models: In-depth analysis on the abstraction and reasoning corpus," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 6, pp. 1–52, 2025.

[87] Y. Sakai, H. Kamigaito, and T. Watanabe, "Revisiting Compositional Generalization Capability of Large Language Models Considering Instruction Following Ability," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 31219–31238. doi: 10.18653/v1/2025.acl-long.1508.

[88] D. M. Schmidt, R. Schubert, and P. Cimiano, "CompoST: A Benchmark for Analyzing the Ability of LLMs to Compositionally Interpret Questions in a QALD Setting," in *The Semantic Web – ISWC 2025*, Cham: Springer Nature Switzerland, 2026, pp. 3–22. doi: 10.1007/978-3-032-09527-5_1.

[89] R. Kumon and H. Yanaka, "Analyzing the Inner Workings of Transformers in Compositional Generalization," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 8529–8540. doi: 10.18653/v1/2025.naacl-long.432.

[90] Z. Xu, Z. Yang, and H. Wang, "MC^2: A Minimum-Coverage and Dataset-Agnostic Framework for Compositional Generalization of LLMs on Semantic Parsing," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 7694–7706. doi: 10.18653/v1/2025.findings-emnlp.406.

[91] T. Vrabcová and P. Sojka, "Negation Disrupts Compositionality in Language Models: The Czech Usecase," *The Eighteenth Workshop on Recent Advances in Slavonic Natural Language Processing*, pp. 17–24, 2024.

[92] A. Moisio, M. Creutz, and M. Kurimo, "Evaluating Morphological Generalisation in Machine Translation by Distribution-Based Compositionality Assessment," in *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, Tórshavn, Faroe Islands: University of Tartu Library, May 2023, pp. 738–751. [Online]. Available: https://aclanthology.org/2023.nodalida-1.75/

[93] A. Moisio, M. Creutz, and M. Kurimo, "On using distribution-based compositionality assessment to evaluate compositional generalisation in machine translation," in *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, Singapore: Association for Computational Linguistics, Dec. 2023, pp. 204–213. doi: 10.18653/v1/2023.genbench-1.17.

[94] Z. Wang and D. Hershcovich, "On Evaluating Multilingual Compositional Generalization with Translated Datasets," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada:

Association for Computational Linguistics, Jul. 2023, pp. 1669–1687. doi: 10.18653/v1/2023.acl-long.93.

[95] T. Schultz, “Globalphone: A multilingual speech and text database developed at Karlsruhe University,” in *Proceedings of ICSLP 2002*, 2002, pp. 345–348. doi: 10.21437/ICSLP.2002-151.

[96] T. Schultz and A. Waibel, “Language independent and language adaptive large vocabulary speech recognition,” in *Proceedings of ICSLP 1998*, 1998, p. paper 0577. doi: 10.21437/ICSLP.1998-751.

[97] H. Höge, C. Draxler, H. van den Heuvel, F. T. Johansen, E. Sanders, and H. S. Tropf, “Speechdat multilingual speech databases for teleservices: across the finish line,” in *Proceedings of Eurospeech 1999*, 1999, pp. 2699–2702. doi: 10.21437/Eurospeech.1999-578.

[98] A. Iskra, B. Petek, and T. Brøndsted, “Recognition of slovenian speech: within and cross-language experiments on monophones using the speechdat(II),” in *Proceedings of Eurospeech 2001*, 2001, pp. 2777–2780. doi: 10.21437/Eurospeech.2001-650.

[99] Z. Kačič, B. Horvat, and A. Zögling, “Issues in Design and Collection of Large Telephone Speech Corpus for Slovenian Language,” in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC’00)*, Athens, Greece: European Language Resources Association (ELRA), May 2000. [Online]. Available: https://aclanthology.org/L00-1185/

[100] A. Žgank, Z. Kačič, and B. Horvat, “Preliminary Evaluation of Slovenian Mobile Database PoliDat,” in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC’02)*, Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA), May 2002. [Online]. Available: https://aclanthology.org/L02-1089/

[101] A. Vandecatseye *et al.*, “The COST278 Pan-European Broadcast News Database,” in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, Lisbon, Portugal: European Language Resources Association (ELRA), May 2004. [Online]. Available: https://aclanthology.org/L04-1055/

[102] R. Ardila *et al.*, “Common Voice: A Massively-Multilingual Speech Corpus,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520/

[103] N. Ljubešić, D. Koržinek, P. Rupnik, and I.-P. Jazbec, “ParlaSpeech-HR - a Freely Available ASR Dataset for Croatian Bootstrapped from the ParlaMint Corpus,” in *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, Jun. 2022, pp. 111–116. [Online]. Available: https://aclanthology.org/2022.parlaclarin-1.16/

[104] T. Erjavec *et al.*, “The ParlaMint corpora of parliamentary proceedings,” *Lang Resources & Evaluation*, vol. 57, no. 1, pp. 415–448, Mar. 2023, doi: 10.1007/s10579-021-09574-0.

[105] A. Zgank, D. Verdonik, A. Z. Markus, and Z. Kacic, “BNSI Slovenian broadcast news database - speech and text corpus,” in *Proceedings of Interspeech 2005*, 2005, pp. 1537–1540. doi: 10.21437/Interspeech.2005-451.

[106] A. Zgank, T. Rotovnik, M. Grasic, M. Kos, D. Vlaj, and Z. Kacic, “Sloparl - slovenian parliamentary speech and text corpus for large vocabulary continuous speech

recognition," in *Proceedings of Interspeech 2006*, 2006, p. paper 1493. doi: 10.21437/Interspeech.2006-50.

[107] A. Žgank, M. S. Maučec, and D. Verdonik, "The SI TEDx-UM speech database: a new Slovenian Spoken Language Resource," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 4670–4673. [Online]. Available: https://aclanthology.org/L16-1740/

[108] D. Verdonik, K. Dobrovoljc, T. Erjavec, and N. Ljubešić, "Gos 2: A New Reference Corpus of Spoken Slovenian," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia: ELRA and ICCL, May 2024, pp. 7825–7830. [Online]. Available: https://aclanthology.org/2024.lrec-main.691/

[109] D. Verdonik *et al.*, "Strategies for managing time and costs in speech corpus creation: insights from the Slovenian ARTUR corpus," *Lang Resources & Evaluation*, vol. 59, no. 3, pp. 1899–1924, Sep. 2025, doi: 10.1007/s10579-024-09792-2.

[110] S. Martincic-Ipsic and I. Ipsic, "VEPRAD: a Croatian speech database of weather forecasts," in *Proceedings of the 25th International Conference on Information Technology Interfaces, 2003. ITI 2003.*, Jun. 2003, pp. 321–326. doi: 10.1109/ITI.2003.1225364.

[111] A. Žgank, D. Verdonik, A. Z. Markuš, and Z. Kačič, "SINOD - Slovenian non-native speech database," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy: European Language Resources Association (ELRA), May 2006. [Online]. Available: https://aclanthology.org/L06-1085/

[112] B. Marković, S. T. Jovičić, J. Galić, and Đ. Grozdić, "Whispered Speech Database: Design, Processing and Application," in *Text, Speech, and Dialogue*, Berlin, Heidelberg: Springer, 2013, pp. 591–598. doi: 10.1007/978-3-642-40585-3_74.

[113] S. Ostrogonac, S. Suzić, M. Bojanić, and E. Pakoci, "Speech resources for a Serbian LVCSR system," in *2013 21st Telecommunications Forum Telfor*, Nov. 2013, pp. 478–481. doi: 10.1109/TELFOR.2013.6716271.

[114] V. Delić *et al.*, "Speech and Language Resources within Speech Recognition and Synthesis Systems for Serbian and Kindred South Slavic Languages," in *Proceedings of the 15th International Conference on Speech and Computer*, in SPECOM 2013, vol. 8113. Berlin, Heidelberg: Springer-Verlag, Sep. 2013, pp. 319–326. doi: 10.1007/978-3-319-01931-4_42.

[115] F. Mihelič, J. Gros, E. Nöth, and V. Warnke, "Labeling of Prosodic Events in Slovenian Speech Database GOPOLIS," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece: European Language Resources Association (ELRA), May 2000. [Online]. Available: https://aclanthology.org/L00-1220/

[116] M. Rojc, Z. Kačič, and D. Verdonik, "Design and Implementation of the Slovenian Phonetic and Morphology Lexicons for the Use in Spoken Language Applications," in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, European Language Resources Association (ELRA), May 2002. doi: 10.63317/4ywo9e4nrvmq.

[117] M. Rojc and Z. Kacic, "Efficient Development of Lexical Language Resources and their Representation," *International Journal of Speech Technology*, vol. 6, pp. 259–275, Jun. 2003, doi: 10.1023/A:1023418220679.

[118] U. Ziegenhain, H. Fersoe, H. van den Heuvel, and A. Moreno, “LC-STAR II: Starring more Lexica,” in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco: European Language Resources Association (ELRA), May 2008. [Online]. Available: https://aclanthology.org/L08-1486/

[119] J. Ž. Gros, V. Cvetko-Orešnik, P. Jakopin, and A. Mihelič, “SI-PRON: A Pronunciation Lexicon for Slovenian,” in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy: European Language Resources Association (ELRA), May 2006. [Online]. Available: https://aclanthology.org/L06-1057/

[120] K. Dobrovoljc, “Treebanking Spoken Slovenian: New Data, Models, and Lessons Learned,” *Contributions to Contemporary History*, vol. 65, no. 3, Dec. 2025, doi: 10.51663/pnz.65.3.01.

[121] B. Dropuljić, M. T. Chmura, A. Kolak, and D. Petrinović, “Emotional speech corpus of Croatian language,” in *2011 7th International Symposium on Image and Signal Processing and Analysis (ISPA)*, Sep. 2011, pp. 95–100. [Online]. Available: https://ieeexplore.ieee.org/document/6046587

[122] T. Justin, V. Štruc, J. Žibert, and F. Mihelič, “Development and Evaluation of the Emotional Slovenian Speech Database - EmoLUKS,” in *Text, Speech, and Dialogue*, Cham: Springer International Publishing, 2015, pp. 351–359. doi: 10.1007/978-3-319-24033-6_40.

[123] J. Kuvač Kraljević and G. Hržica, “Croatian adult spoken language corpus (HrAL),” *FLUMINENSIA : časopis za filološka istraživanja*, vol. 28, no. 2, pp. 87–102, 2016.

[124] I. Mlakar, D. Verdonik, S. Majhenič, and M. Rojc, “Understanding conversational interaction in multiparty conversations: the EVA Corpus,” *Lang Resources & Evaluation*, vol. 57, no. 2, pp. 641–671, Jun. 2023, doi: 10.1007/s10579-022-09627-y.

[125] P. Wasserscheidt *et al.*, “Corpus-based analysis of spoken narratives. Introducing a corpus and a search tool,” *Govor*, vol. 37, no. 2, pp. 149–178, 2020, doi: 10.22210/govor.2020.37.08.

[126] N. Poropat Jeletić, G. Hržica, and E. Moscarda Mirković, “The Corpus of Spoken Istrovenetian/Fiuman and Croatian (C-ORAL-IC),” *FLUMINENSIA : časopis za filološka istraživanja*, vol. 37, no. 1, pp. 89–109, Jul. 2025, doi: 10.31820/f.37.1.4.

[127] B. Petek, “Towards voice-interactive telephone services in Slovenia: on prosody of digits using the sociolinguistic framework,” in *Proceedings of Eurospeech 1995*, 1995, pp. 1015–1018. doi: 10.21437/Eurospeech.1995-250.

[128] B. Imperl, “Multilingual connected digits and natural numbers recognition in the telephone speech dialog systems,” in *ISIE ’99. Proceedings of the IEEE International Symposium on Industrial Electronics (Cat. No.99TH8465)*, Jul. 1999, pp. 188–192 vol.1. doi: 10.1109/ISIE.1999.801782.

[129] A. Žgank *et al.*, “The COST 278 MASPER Initiative - Crosslingual Speech Recognition with Large Telephone Databases,” in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, Lisbon, Portugal: European Language Resources Association (ELRA), May 2004. [Online]. Available: https://aclanthology.org/L04-1090/

[130] B. Kotnik, Z. Kačič, and B. Horvat, “The Development and Integration of the LDA-Toolkit Into COST249 SpeechDat(II) SIG Reference Recognizer,” in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, Lisbon, Portugal: European Language Resources Association (ELRA), May 2004. [Online]. Available: https://aclanthology.org/L04-1047/

[131] D. Vlaj, B. Kotnik, Z. Kacic, and B. Horvat, “Usage of frame dropping and frame attenuation algorithms in automatic speech recognition systems,” in *The IEEE Region 8 EUROCON 2003. Computer as a Tool.*, Sep. 2003, pp. 149–152 vol.2. doi: 10.1109/EURCON.2003.1248170.

[132] A. Zgank and Z. Kacic, “Conversion from phoneme based to grapheme based acoustic models for speech recognition,” in *Proceedings of Interspeech 2006*, 2006, p. paper 1500. doi: 10.21437/Interspeech.2006-444.

[133] J. Zibert, S. Martincic-Ipsic, I. Ipsic, and F. Mihelic, “Bilingual speech recognition of Slovenian and Croatian weather forecasts,” in *Proceedings EC-VIP-MC 2003. 4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications*, Jul. 2003, pp. 637–642 vol.2. doi: 10.1109/VIPMC.2003.1220535.

[134] M. Gulić, D. Lučanin, and A. Šimić, “A digit and spelling speech recognition system for the Croatian language,” in *2011 Proceedings of the 34th International Convention MIPRO*, May 2011, pp. 1673–1678. [Online]. Available: https://ieeexplore.ieee.org/document/5967330

[135] M. S. Maučec, T. Rotovnik, and M. Zemljak, “Modelling Highly Inflected Slovenian Language,” *International Journal of Speech Technology*, vol. 6, no. 3, pp. 245–257, Jul. 2003, doi: 10.1023/A:1023466103841.

[136] P. Geutner, M. Finke, and A. Waibel, “Selection criteria for hypothesis driven lexical adaptation,” in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, Mar. 1999, pp. 617–620 vol.2. doi: 10.1109/ICASSP.1999.759742.

[137] A. Waibel, P. Geutner, L. M. Tomokiyo, T. Schultz, and M. Woszczyna, “Multilinguality in speech and spoken language systems,” *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1297–1313, Aug. 2000, doi: 10.1109/5.880085.

[138] S. Martinčić-Ipšić, S. Ribarić, and I. Ipšić, “Acoustic Modelling for Croatian Speech Recognition and Synthesis,” *Informatica*, vol. 19, no. 2, pp. 227–254, Apr. 2008.

[139] S. Martinčić-Ipšić, M. Pobar, and I. Ipšić, “Croatian Large Vocabulary Automatic Speech Recognition,” *Automatika*, vol. 52, no. 2, pp. 147–157, Jan. 2011, doi: 10.1080/00051144.2011.11828413.

[140] B. Dropuljić and D. Petrinović, “Development of Acoustic Model for Croatian Language Using HTK,” *Automatika*, vol. 51, no. 1, pp. 79–88, Jan. 2010, doi: 10.1080/00051144.2010.11828357.

[141] N. Jakovljevic and D. Pekar, “Description of Training Procedure for AlfaNum Continuous Speech Recognition System,” in *EUROCON 2005 - The International Conference on “Computer as a Tool,”* Nov. 2005, pp. 1646–1649. doi: 10.1109/EURCON.2005.1630286.

[142] I. Ipsic, “Acoustic and language modelling for spoken dialog systems,” in *ISPA 2001. Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis. In conjunction with 23rd International Conference on Information Technology Interfaces (IEEE Cat.*, Jun. 2001, pp. 441–444. doi: 10.1109/ISPA.2001.938670.

[143] A. Zgank, Z. Kacic, and B. Horvat, “Data driven generation of broad classes for decision tree construction in acoustic modeling,” in *Proceedings of Eurospeech 2003*, 2003, pp. 2505–2508. doi: 10.21437/Eurospeech.2003-687.

[144] N. Jakovljevic, D. Miskovic, M. Janev, M. Secujski, and V. Delic, “Comparison of Linear Discriminant Analysis Approaches in Automatic Speech Recognition,” *Elektronika ir Elektrotechnika*, vol. 19, no. 7, pp. 76–79, Sep. 2013, doi: 10.5755/j01.eee.19.7.5167.

[145] T. Rotovnik, M. S. Maucec, B. Horvat, and Z. Kacic, “A comparison of HTK, ISIP and julius in slovenian large vocabulary continuous speech recognition,” in *Proceedings of ICSLP 2002*, 2002, pp. 681–684. doi: 10.21437/ICSLP.2002-227.

[146] N. M. Jakovljević, D. Mišković, E. Pakoci, T. Grbić, and V. Delić, “A comparison of several variants of GMM on speech recognition task,” in *2013 21st Telecommunications Forum Telfor*, Nov. 2013, pp. 466–469. doi: 10.1109/TELFOR.2013.6716268.

[147] N. T. Vu, F. Kraus, and T. Schultz, “Multilingual a-stabil: A new confidence score for multilingual unsupervised training,” in *2010 IEEE Spoken Language Technology Workshop*, Dec. 2010, pp. 183–188. doi: 10.1109/SLT.2010.5700848.

[148] F. Diehl, A. Moreno, and E. Monte, “Crosslingual acoustic model development for automatic speech recognition,” in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, Dec. 2007, pp. 425–430. doi: 10.1109/ASRU.2007.4430150.

[149] F. Stahlberg, T. Schlippe, S. Vogel, and T. Schultz, “Towards automatic speech recognition without pronunciation dictionary, transcribed speech and text resources in the target language using cross-lingual word-to-phoneme alignment,” in *Proceedings of SLTU 2014*, 2014, pp. 73–80. [Online]. Available: https://www.isca-archive.org/sltu_2014/stahlberg14_sltu.html

[150] A. Zgank, B. Imperl, F. T. Johansen, Z. Kacic, and B. Horvat, “Crosslingual speech recognition with multilingual acoustic models based on agglomerative and tree-based triphone clustering,” in *Proceedings of Eurospeech 2001*, 2001, pp. 2725–2729. doi: 10.21437/Eurospeech.2001-637.

[151] J. Nouza, P. Cerva, and M. Kucharova, “Cost-Efficient Development of Acoustic Models for Speech Recognition of Related Languages,” *Radioengineering*, vol. 22, no. 3, pp. 866–873, Sep. 2013.

[152] A. Žgank, T. Rotovnik, and M. S. Maučec, “Slovenian spontaneous speech recognition and acoustic modeling of filled pauses and onomatopoeas,” *WSEAS Trans. Sig. Proc.*, vol. 4, no. 7, pp. 388–397, Jul. 2008.

[153] M. Kos, M. Rojc, A. Žgank, Z. Kačič, and D. Vlaj, “A speech-based distributed architecture platform for an intelligent ambience,” *Computers & Electrical Engineering*, vol. 71, pp. 818–832, Oct. 2018, doi: 10.1016/j.compeleceng.2017.07.010.

[154] B. Kotnik, D. Vlaj, and B. Horvat, “Efficient Noise Robust Feature Extraction Algorithms for Distributed Speech Recognition (DSR) Systems,” *International Journal of Speech Technology*, vol. 6, no. 3, pp. 205–219, Jul. 2003, doi: 10.1023/A:1023410018862.

[155] R. Modic, B. Lindberg, and B. Petek, “Comparative wavelet and MFCC speech recognition experiments on the Slovenian and English speechdat2,” in *Proceedings of NOLISP 2003*, 2003, p. paper 016. [Online]. Available: https://www.isca-archive.org/nolisp_2003/modic03_nolisp.html

[156] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, “Development of the RWTH transcription system for slovenian,” in *Proceedings of Interspeech 2013*, 2013, pp. 3107–3111. doi: 10.21437/Interspeech.2013-677.

[157] B. Popović, S. Ostrogonac, E. Pakoci, N. Jakovljević, and V. Delić, “Deep Neural Network Based Continuous Speech Recognition for Serbian Using the Kaldi Toolkit,” in *Speech and Computer*, Cham: Springer International Publishing, 2015, pp. 186–192. doi: 10.1007/978-3-319-23132-7_23.

[158] B. Popović, E. Pakoci, N. Jakovljević, G. Kočiš, and D. Pekar, “Voice assistant application for the Serbian language,” in *2015 23rd Telecommunications Forum Telfor*, Nov. 2015, pp. 858–861. doi: 10.1109/TELFOR.2015.7377600.

[159] E. Pakoci, B. Popović, and D. J. Pekar, “Improvements in Serbian Speech Recognition Using Sequence-Trained Deep Neural Networks,” *Informatics and Automation*, vol. 3, no. 58, pp. 53–76, Jun. 2018, doi: 10.15622/sp.58.3.

[160] J. Nouza, R. Safarik, and P. Cerva, “ASR for South Slavic Languages Developed in Almost Automated Way,” in *Proceedings of Interspeech 2016*, 2016, pp. 3868–3872. doi: 10.21437/Interspeech.2016-747.

[161] A. Zgank and M. S. Maucec, “Influence of Emotional Speech on Continuous Speech Recognition,” in *2020 ELEKTRO*, May 2020, pp. 1–4. doi: 10.1109/ELEKTRO49696.2020.9130316.

[162] E. Pakoci, B. Popović, and D. Pekar, “Using Morphological Data in Language Modeling for Serbian Large Vocabulary Speech Recognition,” *Computational Intelligence and Neuroscience*, vol. 2019, no. 1, p. 5072918, 2019, doi: 10.1155/2019/5072918.

[163] E. T. Pakoci and B. Z. Popović, “Recurrent Neural Networks and Morphological Features in Language Modeling for Serbian,” in *2021 29th Telecommunications Forum (TELFOR)*, Nov. 2021, pp. 1–8. doi: 10.1109/TELFOR52709.2021.9653410.

[164] B. Popović, E. Pakoci, and D. Pekar, “Transfer Learning in Automatic Speech Recognition for Serbian,” in *2019 27th Telecommunications Forum (TELFOR)*, Nov. 2019, pp. 1–4. doi: 10.1109/TELFOR48224.2019.8971335.

[165] B. Z. Popović, E. T. Pakoci, and D. J. Pekar, “Transfer learning for domain and environment adaptation in Serbian ASR,” *Telfor Journal*, vol. 12, no. 2, pp. 110–115, 2020, doi: 10.5937/telfor2002110P.

[166] N. Hailu, I. Siegert, and A. Nürnberger, “Improving Automatic Speech Recognition Utilizing Audio-codecs for Data Augmentation,” in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, Sep. 2020, pp. 1–5. doi: 10.1109/MMSP48831.2020.9287127.

[167] H. Fazlić, A. A. Almisreb, and N. M. Tahir, “Deep Learning-Based Audio-Visual Speech Recognition for Bosnian Digits,” *JKUKM*, vol. 36, no. 1, pp. 147–154, Jan. 2024, doi: 10.17576/jkukm-2024-36(1)-14.

[168] A. Žgank and G. Donaj, “Slovenian End-to-End Automatic Speech Recognition and the Impact of Emotional Speech,” in *2025 32nd International Conference on Systems, Signals and Image Processing (IWSSIP)*, Skopje, North Macedonia: IEEE, Jun. 2025, pp. 1–5. doi: 10.1109/IWSSIP66997.2025.11151959.

[169] M. Sepesy Maučec, D. Verdonik, and G. Donaj, “Sequence-to-Sequence Models and Their Evaluation for Spoken Language Normalization of Slovenian,” *Applied Sciences*, vol. 14, no. 20, p. 9515, Oct. 2024, doi: 10.3390/app14209515.

[170] S. Zhang, J. Hale, M. Renwick, Z. Vrzić, and K. Langston, “An Evaluation of Croatian ASR Models for Čakavian Transcription,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia: ELRA and ICCL, May 2024, pp. 1098–1104. [Online]. Available: https://aclanthology.org/2024.lrec-main.98/

[171] A. Zgank, “Influence of Highly Inflected Word Forms and Acoustic Background on the Robustness of Automatic Speech Recognition for Human–Computer Interaction,” *Mathematics*, vol. 10, no. 5, p. 711, Jan. 2022, doi: 10.3390/math10050711.

[172] N. Mikelić Preradović and L. Nacinovic Prskalo, “System for Automatic Assignment of Lexical Stress in Croatian,” *Electronics*, vol. 11, no. 22, p. 3687, Jan. 2022, doi: 10.3390/electronics11223687.

[173] A. Žgank, “Analyzing the influence of narrow-band channel on Slovenian broadcast news speech recognition,” in *2016 ELEKTRO*, May 2016, pp. 116–119. doi: 10.1109/ELEKTRO.2016.7512047.

[174] S. Dobrišek, B. Vesnicer, and F. Mihelič, “A sequential minimization algorithm for finite-state pronunciation lexicon models,” in *Proceedings of Interspeech 2009*, 2009, pp. 720–723. doi: 10.21437/Interspeech.2009-246.

[175] E. Pakoci, D. Pekar, B. Popović, M. Sečujski, and V. Delić, “Overcoming Data Sparsity in Automatic Transcription of Dictated Medical Findings,” in *2022 30th European Signal Processing Conference (EUSIPCO)*, Aug. 2022, pp. 454–458. doi: 10.23919/EUSIPCO55093.2022.9909893.

[176] N. Ljubešić and D. Lauc, “BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian,” in *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, Kyiv, Ukraine: Association for Computational Linguistics, Apr. 2021, pp. 37–42. [Online]. Available: https://aclanthology.org/2021.bsnlp-1.5/

[177] J. Galic and Đ. Grozdić, “Exploring the Impact of Data Augmentation Techniques on Automatic Speech Recognition System Development: A Comparative Study,” *Advances in Electrical and Computer Engineering*, vol. 23, pp. 3–12, Jan. 2023, doi: 10.4316/AECE.2023.03001.

[178] N. Ljubešić, I. Porupski, and P. Rupnik, “Identifying Filled Pauses in Speech Across South and West Slavic Languages,” in *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 1–8. doi: 10.18653/v1/2025.bsnlp-1.1.

[179] A. Žgank and G. Donaj, “Slovenian End-to-End Automatic Speech Recognition and the Impact of Emotional Speech,” in *2025 32nd International Conference on Systems, Signals and Image Processing (IWSSIP)*, Skopje, North Macedonia: IEEE, Jun. 2025, pp. 1–5. doi: 10.1109/IWSSIP66997.2025.11151959.

ABBREVIATION LIST

AI	artificial intelligence
ASR	automatic speech recognition
CER	character error rate
CFQ	Compositional Freebase Questions
CG	compositional generalization
CNN	convolutional neural network
COGS	Compositional Generalization Challenge based on Semantic Interpretation
CTC	Connectionist Temporal Classification
DBCA	Distribution-Based Compositionality Assessment
DNN	deep neural network
E2E	end-to-end
GMM	Gaussian Mixture Model
HDLA	Hypothesis Driven Lexical Adaptation
HLDA	Heteroscedastic Linear Discriminant Analysis
HMM	Hidden Markov Model
HTK	HMM Toolkit
KBQA	Knowledge-Based Question Answering
LLM	large language model
LSTM	long short-term memory
LVCSR	large vocabulary continuous speech recognition
MCD	maximum compound divergence
MFCC	Mel-Frequency Cepstral Coefficient
MRL	morphologically rich language
NLP	natural language processing
OOD	out-of-distribution
OOV	out-of-vocabulary
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses

RNN	recurrent neural network
SCAN	Simplified version of the CommAI Navigation task
SLU	spoken language understanding
WER	word error rate
XAI	explainable artificial intelligence

ABSTRACT

Human language operates on the principle of compositionality—the ability to understand and create novel statements by recombining familiar words and structures. In artificial intelligence, this is known as compositional generalization (CG). Despite the success of modern sequence-to- sequence models and large language models (LLMs), recent evidence shows they often struggle to systematically generalize to new combinations of familiar elements, relying heavily on distributional patterns instead. Most research on CG has been restricted to English text, largely ignoring spoken language technologies and morphologically rich languages (MRLs) such as Croatian and related South Slavic languages.

To address this gap, this qualification exam paper presents a systematic PRISMA-based literature review mapping two distinct domains: compositional generalization in natural language processing (76 included studies) and automatic speech recognition (ASR) for South Slavic languages (86 included studies). The review reveals a complete lack of intersection between these fields. The CG literature indicates that merely increasing model scale does not resolve systematicity deficits, highlighting the need for structural inductive biases. Concurrently, the South Slavic ASR literature demonstrates continuous technological progress, but identifies morphological richness and high out-of-vocabulary (OOV) rates as persistent challenges. Current ASR systems handle this morphosyntactic complexity empirically, but they are evaluated almost exclusively through aggregate metrics like Word Error Rate (WER) and Character Error Rate (CER). These metrics mask a critical vulnerability: systems may simply memorize frequent word forms without learning the underlying compositional rules of inflection.

The synthesis of these findings forms the motivation for the upcoming doctoral research. The thesis will propose integrating diagnostic, morphology-aware compositional evaluation frameworks into South Slavic ASR systems. Specifically, numeral expressions will serve as the initial testing ground, as they combine productively from a well-defined set of primitives under strict morphological rules. This methodology will allow for the precise measurement of whether a speech model genuinely understands compositional structure or merely relies on frequency-based memorization within morphologically rich environments.

SAŽETAK

Ljudski jezik funkcionira prema načelu kompozicionalnosti – sposobnosti razumijevanja i stvaranja novih izjava kombiniranjem poznatih riječi i struktura. U umjetnoj inteligenciji to se naziva kompozicijskom generalizacijom (CG). Unatoč uspjehu modernih neuronskih modela i velikih jezičnih modela (LLM), noviji dokazi pokazuju da se oni često bore sa sustavnom generalizacijom na nove kombinacije poznatih elemenata, oslanjajući se prvenstveno na distribucijske obrasce iz treninga. Većina istraživanja CG-a ograničena je na engleski tekst, zanemarujući govorne tehnologije i morfološki bogate jezike poput hrvatskog i srodnih južnoslavenskih jezika.

Kako bi se premostio taj jaz, ovaj rad predstavlja sustavni pregled literature temeljen na PRISMA smjernicama, koji mapira dvije odvojene domene: kompozicijsku generalizaciju u obradi prirodnog jezika (76 uključenih studija) i automatsko prepoznavanje govora (ASR) za južnoslavenske jezike (86 uključenih studija). Pregled ukazuje na potpuni izostanak presjeka između ovih područja. Literatura o CG-u pokazuje da samo povećanje veličine modela ne rješava problem generalizacije, ističući potrebu za strukturnim promjenama u arhitekturama. Istovremeno, literatura o južnoslavenskom ASR-u bilježi stalan tehnološki napredak, ali prepoznaje morfološko bogatstvo i visoku stopu nepoznatih riječi (OOV) kao trajne prepreke. Postojeći ASR sustavi prilagođavaju se toj složenosti, no evaluiraju se gotovo isključivo putem ukupnih metrika pogreške (WER i CER). Takav pristup maskira ključnu ranjivost: sustavi možda samo pamte česte oblike riječi bez stvarnog usvajanja kompozicijskih pravila za promjenu oblika riječi.

Sinteza ovih nalaza predstavlja motivaciju za predstojeće doktorsko istraživanje. Cilj budućeg istraživanja je integracija dijagnostičkih, morfološki svjesnih okvira za evaluaciju kompozicionalnosti u ASR sustave za južnoslavenske jezike. Kao početni evaluacijski testni slučaj predlažu se brojevni izrazi, s obzirom na to da se oni produktivno tvore iz usko definiranog skupa primitiva prema strogim morfološkim pravilima. Ovakav pristup omogućit će precizno mjerenje sposobnosti govornog modela da istinski razumije kompozicijsku strukturu i prilagodi se morfološki bogatom okruženju, za razliku od izravnog pamćenja oslonjenog na učestalost u podacima za treniranje.