


Winograd schemata and other datasets for anaphora resolution in Hungarian

NOÉMI VADÁSZ*  and NOÉMI LIGETI-NAGY

Language Technology Research Group, Hungarian Research Centre for Linguistics, Hungary

Received: April 21, 2022 • Accepted: October 1, 2022

Published online: November 22, 2022

© 2022 The Author(s)



ABSTRACT

The Winograd Schema Challenge (WSC, proposed by Levesque, Davis & Morgenstern 2012) is considered to be the novel Turing Test to examine machine intelligence. Winograd schema questions require the resolution of anaphora with the help of world knowledge and commonsense reasoning. Anaphora resolution is itself an important and difficult issue in natural language processing, therefore, many other datasets have been created to address this issue. In this paper we look into the Winograd schemata and other Winograd-like datasets and the translations of the schemata to other languages, such as Chinese, French and Portuguese. We present the Hungarian translation of the original Winograd schemata and a parallel corpus of all the translations of the schemata currently available. We also adapted some other anaphora resolution datasets to Hungarian. We aim to discuss the challenges we faced during the translation/adaptation process.

KEYWORDS

Winograd schema, anaphora resolution, commonsense reasoning, dataset, corpus

1. INTRODUCTION

Based on the classical definition of Halliday et al. (1976) anaphora is cohesion (presupposition) which points back to a previous item. The “pointing back” item is called an anaphor and the element to which it refers is its antecedent. The process of determining the antecedent of an anaphor is called anaphora resolution. Coreference corpora, Winograd schemata and other

* Corresponding author. E-mail: vadasz.noemi@nytud.hu

similar datasets can be used for training and testing approaches to anaphora resolution. In this paper we dive into the world of these datasets concentrating on the problem of pronominal anaphora.

2. BACKGROUND

In this section, after describing the original idea of Winograd schemata and how it is used for testing whether the computer can resolve pronominal anaphora, we turn to Hungarian coreference corpora as they form an important milestone in the field of coreference and anaphora resolution for Hungarian.

2.1. The Winograd schemata

As [Hirst \(1981\)](#) in one of the ground-breaking studies of automatic anaphora resolution has illustrated the phenomenon with sentence pairs, Winograd schemata are also sentence pairs where the contents of the two sentences are as similar as possible (differing in one word or phrase) and the target pronouns are identical lexically, but they refer back to different antecedents. The idea behind the Winograd schemata is based on that in these problems grammatical behavior (e.g. binding) is not enough to find the antecedent of a pronoun, world knowledge and commonsense reasoning are also needed.

The two sentences in Example (1) differ only in the verb, yet the pronominal subject of this verb points back to an other antecedent. In order to connect the pronoun to its antecedent correctly the resolver – be it a human or a model – must know that the city councilmen fear violence, and the demonstrators are the ones who advocate violence (at least according to the city councilmen).

- (1) The city councilmen refused the demonstrators a permit because they [feared/advocated] violence.
Who [feared/advocated] violence?
a. The city councilmen
b. The demonstrators

2.2. The Winograd schema challenge as the new Turing test

[Turing \(1950\)](#) proposed a game as a way to test machine's intelligence. Turing himself rejects to discuss the question whether a machine can think – as it is meaningless –; instead, he proposes the so-called 'imitation game', and the question being worth discussing is whether a computer can perform well in that game.

The imitation game is the following (as proposed in [Turing 1950](#)): an interrogator is asking questions, while a computer and a human are answering those. The interrogator does not see the machine and the person but is aware of the fact that a computer and a human are answering. The goal of the machine is to cause the interrogator to falsely conclude that the computer is the human participant. Turing predicted that by the end of the 20th century the interrogators would



only have 70% chance of making the right identification after 5 min of questioning (Turing 1950, 442).

The phrase “Turing Test” is also used as a general term for any test aiming to measure computer’s “intelligence”. The imitation game itself provoked a long lasting debate on how to define “thinking” and “intelligence” in the field of artificial intelligence and many papers argued that Turing’s test is far from suitable for measuring the intelligence of machines (a few recent summaries on the topic; Copeland 2000; Damassino 2020; Neufeld & Finnstad 2020).

Levesque et al. (2012) proposes a set of Winograd schemata as a novel test for AI programs, along the lines of the Turing Test. A Winograd schema has to meet three criteria to be involved into the challenge:

1. it has to be easily disambiguated by a human reader
2. it must not be solveable by selectional restrictions
3. it must be not googleable

The advantage of this novel challenge is that it is straightforward: the answer to the schemata is a binary choice. Moreover, it is expressive: any non-experts can determine that a program that fails to get the correct answer is not “intelligent” enough; that program is far from human understanding. Finally, the schemata are difficult: anaphora resolution is an obvious task for a human, but still hard for the state-of-the-art algorithms. The reason for this is that only world knowledge and reasoning can help solve these problems.

The first condition can be easily checked with human annotators, but the other two conditions can be criticized. First, the dataset is already on the web together with the labels in several languages, so for every schema the solution itself is googleable. In addition to selection restrictions, it has been shown that the task can often be solved by association (for details, see Section 3.4.2).

2.3. Hungarian coreference corpora

Before we move on to Winograd schemata and other datasets, we need to look at two important resources for coreference resolution in Hungarian.

SzegedKoref (Vincze et al. 2018) was compiled from a part of Szeged Korpusz (Csendes et al. 2005) consisting of student essays and newspaper articles, altogether 55,763 tokens. 2,456 coreference chains are annotated in the corpus, of which 1,851 are pronominal anaphora. It is available for research and educational purposes on request.

SzegedKoref inspired an other coreference corpus, KorKor (Vadász 2020), which also contains anaphoric and coreference relations. The corpus consists of Wikipedia and newspaper articles, altogether 31,492 tokens including punctuation and zero elements (zero substantives, ellipted verbs and pronouns). 2,015 pronouns (of 9 categories) are marked and disambiguated in the texts. KorKor corpus is freely available.

Both corpora contain manually corrected morphological tags as well as syntactic annotation. The importance of these resources is also reflected in the fact that both contain zero pronouns for dropped subjects, objects and possessors. These are unique resources in this regard, because they allow the phenomenon – namely pronominal anaphora with dropped pronouns – to be studied on living texts. It is not a negligible topic for pro-drop languages, because for instance in KorKor corpus three-quarters of the pronouns are dropped.



The purpose of the coreference corpora differs from Winograd schemata. Coreference corpora are built for training or testing coreference resolution approaches, and since coreference relation (e.g. synonym, repetition) can span through sentences, longer, coherent texts are needed for capturing these phenomena. In most coreference corpora, pronominal anaphora is annotated as a subtype of coreference and resolving anaphoric relations is always included into the task of coreference resolution. Resolving both anaphoric and coreference relations is required for interpreting a text, however the differences between them should be noted. Coreference is a symmetric transitive relation, while anaphora is not, but anaphora is context-dependent. On one hand, resolving coreference needs lexical and semantic knowledge, on the other hand, behind the behavior of pronouns lay syntactic and morphosyntactic rules. However, technically, the task in both cases is clear: finding connecting elements in the text.

As both of SzegedKoref and KorKor are built using a fine-grained tagset for the different subclasses of pronominal anaphora and coreference, they are suitable for anaphora resolution and coreference resolution separately as well.

3. DATASETS

3.1. The use of the datasets

Oxford Dictionary defines a dataset as “a collection of data that is treated as a single unit by a computer”. This means that a dataset contains a lot of separate pieces of data but can be used to train an algorithm with the goal of finding predictable patterns inside the whole dataset. In other words, within the ranks of natural language processing a dataset is a corpus of texts annotated for the specific needs of the given task; what is the aspect of the language or the aspect of its use that we want the algorithm to know or to be able to solve.

A single training dataset that has already been annotated is usually split into several parts, which is needed to check how well the training of the model went. For this purpose, a testing dataset (usually called test set) is usually separated from the training data (train set). Next, a validation dataset (validation set or development set) is used to avoid training the algorithms on the same type of data too long and thus making biased predictions.

A few examples of how Winograd schemata are stored and annotated in datasets used for training machine learning algorithms are shown in this section.

3.2. Benchmark datasets with winograd schemata

In the fields of machine learning and NLP, benchmarks consist of one or more databases, their corresponding metrics and methods of evaluation. Benchmarks provide a standard for measuring performance of various architectures which the professional community agrees upon. For this reason, the most recent databases are based on already existing tasks' corpora (such as GLUE, Wang et al. 2018; or XTREME, Hu et al. 2020) or they are built in accordance of recommendations from the community (e.g., SuperGlue, Wang, Pruksachatkun et al. 2020; or BIG-Bench, Ghazal et al. 2017).

The General Language Understanding Evaluation (GLUE) benchmark was presented in 2019. Corpora were selected in order to provide various domains and difficulty levels for evaluating language comprehension. GLUE consists of nine pre-existing, although slightly modified databases. Winograd schemata are represented here as a natural language inference (NLI) task:



sentence pairs were constructed from the original schemata by replacing the ambiguous pronoun with each possible referent. The task is to predict if the sentence with the pronoun substituted (*sentence2* in Example (2)) is entailed by the original sentence (*sentence1* in Example (2)). The label of a sentence pair is ‘0’, if *sentence2* is not entailed by *sentence1*, and ‘1’ if it is entailed.

- (2) *sentence1*: The drain is clogged with hair. It has to be cleaned.
sentence2: The hair has to be cleaned.
label: 0

The WNLI task in the GLUE benchmark consists of 849 sentence pairs (634 sentence pairs in the train set, 70 sentence pairs in the development set and 145 sentence pairs in the test set) which is more than the 600 sentence pairs one can derive from the original 150 English schemata. The authors state that they use a small evaluation set consisting of new examples derived from fiction books that was shared with them privately by the authors of the original schemata.¹ On top of that, WNLI includes schemata from other pronoun disambiguation tasks as well, see Section 3.4.5.

The creation of another benchmark dataset collection, SuperGLUE was motivated by the fact that GLUE has already proved too easy for language models, thus the authors tried to create corpora suitable for more difficult tasks. SuperGLUE also contains Winograd schemata as binary classification, where each example consists of a sentence with a marked pronoun and noun, and the task is to determine if the pronoun refers to that noun (see Example (3)).

- (3) The large ball crashed right through **the table** because **it** was made of styrofoam.

On the WNLI dataset in GLUE (where the schemata are presented as a natural inference task) Vega v1 (Wang, Xu et al. 2020) has the state-of-the-art result so far, an accuracy of 97.9%. On the WSC dataset of SuperGLUE, ERNIE 3.0 (Sun et al. 2021) and the Microsoft Turing model T-NLRv5 share the state-of-the-art result, 97.3% accuracy. The human baseline for both tasks is 100%.

3.3. Winograd schemata in other languages

Winograd schemata have been adapted to other languages. The web page of the Winograd schemata² presents a Japanese translation of the dataset (in two versions: one with English proper names, and another with Japanese names), but no additional information is available on the translation/adaptation method except for the names of the translators. A Hebrew translation is also linked, but the links do not work.³ There is also a small collection (12 pieces) of schemata translated into Chinese. Russian SuperGLUE (Shavrina et al. 2020) incorporates a collection of Russian Winograd schemata, but the authors do not provide any documentation of the method they applied.

We briefly present the translation of Winograd schemata into French, Portuguese and Mandarin Chinese as they are well documented in academic papers.

¹They say “evaluation set” but may refer to the test set which is distributed without labels.

²<https://cs.nyu.edu/davise/papers/WinogradSchemas/WS.html>

³Date of access: 28/03/2022.



3.3.1. French. Amsili & Seminck (2017) translated the English schemata to French. The collection contains 107 Winograd schemata. The authors thoroughly describe the process of the translation and discuss the challenges they met. The first problem for a French translation was to ensure that the pronoun and the two possible antecedents have the same number and gender. The schema of Example (2) is not translatable because ‘hair’ in French (*cheveux*) is plural, while ‘drain’ (*siphon*) is singular. To solve this conflict the authors replaced ‘hair’ with ‘soap’ (*savon*).

An other problem they enlist is that a literal translation may cause a schema to be ambiguous, as in Example (4). The French equivalent of ‘indiscreet’ is *indiscrète*; and *une personne indiscrète* can be somebody who tries to find out what should stay secret, so a ‘nosy’ person. In this case they changed ‘indiscreet’ *indiscrète* to ‘talkative’ *bavarde*.

(4) Susan knows all about Ann’s personal problems because she is [nosy/indiscreet].

Items that they could not find a solution for were excluded from the final set.

3.3.2. Portuguese. De Melo et al. (2019) introduces the (Brazilian) Portuguese set of Winograd schemata. Similarly to Amsili & Seminck (2017), they initiated their collection from the English set of schemata. Three native Portuguese speakers worked on translating the sentences: each sentence was translated by one of the speakers and validated by the other two. Eight schemata were discarded as no suitable Portuguese translation was found for them. Some of the schemata had to be modified due to similar reasons as those presented in (Amsili & Seminck 2017).

(5) The trophy doesn’t fit into the brown suitcase because it is too large.

In Example (5), the pronoun *ele* ‘it’ can be easily resolved as it refers to a masculine object and the only masculine object in the sentence is *troféu* ‘trophy’ (*maleta* ‘suitcase’ is feminine). These sentences (similarly to the French ones) were modified so that the possible antecedents are of the same gender as the pronoun (in this case, *troféu* was changed to *medalha* ‘medal’, a feminine noun, and the masculine pronoun *ele* was replaced by its feminine counterpart, *ela*).

The Portuguese Winograd schemata are published in two versions. The first version contains the original English names, but in the second version the authors also took care of proper names in the set: famous persons’ names, such as *Shakespeare* or *Madonna* were left as they were; all the other names were replaced by Portuguese names.

3.3.3. Chinese. Bernard & Han (2020) introduce Mandarinograd, the Chinese collection of Winograd schemata. Similarly to the French and Portuguese set of schemata, the Chinese version is also a translation of the English ones. During the translation the problems discussed above (Examples (4) and (5)) were also present in this dataset: for example, the schema *I couldn’t put the pot on the shelf because it was too [high/tall]* could not be applied directly as *tall* (for a pot) and *high* (for a shelf) are both translated as 高. Therefore the word *tall* was replaced by *short*.

3.4. Other WS-like datasets

Additional similar datasets have been created in the pattern of Winograd schemata, which also consist of sentence pairs, or at least aim at resolving ambiguous pronouns, and some contain



more carefully crafted sentences based on critiques of Winograd schemata. In this chapter, we describe resources that are all related to the Winograd schemata in these respects.

3.4.1. The definite pronoun resolution dataset. Rahman & Ng (2012) introduces the notion of *difficult pronouns*. Difficult pronouns appear in such complex cases when the sentences contain two clauses which are separated by a discourse connective and there are two potential antecedent candidates in the first sentence and the definite pronoun in the second sentence – the antecedent of which we are looking for – agrees both antecedent candidates in number, person, gender, semantic class.

According to their findings, otherwise successful anaphora resolution solutions cannot handle difficult pronouns either. One of the basic elements of rule-based solutions, which is based on the principles of syntactic binding theory, does not help here, as the pronoun and candidate are in separate clauses, separated by a discourse connective. The grammatical properties of the pronoun do not help either if they match multiple possible antecedent candidates in number, person, and so on. Traditional machine learning solutions work by exploring the relationships between the training examples seen and surface features, but surface features are worth nothing for difficult pronouns. Thus, in order to find the antecedent of difficult pronouns, it is not enough to thoroughly formalize the linguistic phenomenon or provide many training examples, in fact it requires a deeper understanding of the language.⁴

However, existing anaphora resolution approaches – whether using rule-based, traditional machine learning, or deep-learning techniques – do not even seem to aim to resolve difficult pronouns correctly. Difficult pronouns are rare in standard evaluation corpora (used in Message Understanding Conference (MUC) tasks in 1995 and 1998, ACE (Doddington et al. 2004), OntoNotes (Pradhan et al. 2007)). Instead of dealing with rare but difficult cases, it is enough to aim to solve common and easy-to-use pronoun anaphors, as this can also achieve quite high performance. However, many researchers are concerned about hard cases, and Winograd schemes are explicitly considered complicated, as there is a difficult pronoun in every sentence.

As a reply to the problem that standard evaluation corpora contain only a few example of *difficult pronouns* Rahman & Ng (2012) created a dataset of them as well. The Definite Pronoun Resolution Dataset contains 943 manually created and annotated twin-sentences that meet the criteria of the complex cases. They provided a split to a training and a test set following a 70/30 ratio, therefore the dataset could be used for training and evaluating anaphora resolvers. The data is available⁵ without any restrictions of use.

3.4.2. Winogrande. According to Sakaguchi et al. (2019) the reliability of the Winograd schemata is questionable, because it is suspicious that despite of the hardness of the task (anaphora resolution based on commonsense reasoning) recent advances in neural techniques seem very successful. They presume that the large-scale neural models could exploit certain unwanted biases in the dataset despite the two important criteria of the Winograd schemata aiming to

⁴This is what the latest, so-called deep learning techniques attempt to achieve, in which we can fine-tune models with prior language skills for specific tasks like e.g. anaphora resolution. Datasets like Winograd schemata are required for this fine-tuning step.

⁵<https://www.hlt.utdallas.edu/vince/data/emnlp12/>



avoid biases (i.e. being not googleable and being not solvable by selectional restrictions). Thus, the high performance of these neural models doesn't imply that they managed to solve the problem. There are quite different reasons for their apparent success.

To emphasize the importance of the problem concerning the biases in the schemata Trichelair et al. (2019) showed that 13.5% of the original set of the Winograd schemata is associative, meaning that there is an argument for one antecedent being statistically preferred. In Example (6b) one can easily admit that usually buildings are famous and not maps, but in Example (6a) there is no such a hint. It is not hard to see that a dataset is not the best for evaluating an anaphora resolution system or a language model, if in the case of every seventh to eighth sentence there is a loophole to get around the real problem.

- (6) a. Bill passed the gameboy to John because *his* turn was over.
 b. I'm sure that my map will show this **building**; *it* is very **famous**.

The solution that Sakaguchi et al. (2019) suggests is to take the task out of the hands of the human annotators, because it is difficult for humans to write schemata without accidentally inserting unwanted biases. Their dataset⁶ consists of 44k problems and it was built by a carefully designed crowdsourcing task and an adversarial filtering algorithm for removing bias from the data. While humans can still easily solve the problems trivial (94% accuracy), Winogrande is already really giving up the lesson for the computer, according to their report the best results fall 15–35% below the human performance.

3.4.3. Wino-X. Wino-X consists of cross-lingual and multilingual Winograd schemata in German, French, Russian and English by Emelin & Sennrich (2021). This resource can be used for two distinct purposes: first, to examine the suitability of machine translation for coreference resolution in texts where the task can only be solved with the help of world knowledge and commonsense reasoning, and second, to see how suitable multilingual language models are for commonsense reasoning across languages. Both subsets of Wino-X contain schemata from Winogrande. To avoid problems arising from the translation of the gender of the pronouns only sentences with inanimate antecedent-candidates referred with pronoun *it* have been added to the dataset. The two subsets of Wino-X – MT-Wino-X for evaluating neural machine translation models and LM-Wino-X for multilingual language models – differ in their formats. The former includes the task in the form of a translation test (Example (7)), while the latter adopts the gap-filling format of Winogrande (Example (8)).

- (7) *Source Sentence:* I dusted **the dresser** in the bedroom with **a rag** until *it* was free of dust.
Correct Translation: Ich staubte **die Kommode** im Schlafzimmer mit einem Lappen ab, bis *sie* staubfrei war.
Incorrect Translation: Ich staubte die Kommode im Schlafzimmer mit **einem Lappen** ab, bis *er* staubfrei war.

⁶<https://github.com/allenai/winogrande>



- (8) *EN Context:* Adam chose to sleep on **a sofa** instead of **a bed** because _was much more comfortable.
Correct Filler: the sofa
Incorrect Filler: the bed
DE Context: Adam entschied sich dafür, auf **einem Sofa** statt auf **einem Bett:** zu schlafen, weil _viel bequemer war.
Correct Filler: das Sofa
Incorrect Filler: das Bett

3.4.4. XWINO. XWINO (Tikhonov & Ryabinin 2021) is a multilingual collection of Winograd schemata in six languages that can be used for evaluation of cross-lingual commonsense reasoning capabilities. The multilingual dataset contains pronoun disambiguation problems of six languages. For English they used the original WSC task, the SuperGLUE benchmark, and the Definite Pronoun Resolution Dataset as well. For Portuguese, French, Russian and Japanese they used the translations presented in Section 3.3. Due to the different format of the schemata Mandarinograd was left out.

The dataset⁷ contains 3,961 schemata.

3.4.5. Other pronoun disambiguation problems (PDPs). Anaphora resolution approaches could not only be tested on Winograd schemata. Originally, the first round test set of the 2016 Winograd Schema Challenge was a PDP (Pronoun Disambiguation Problems) test⁸ (Morgens-tern et al. 2016; Davis et al. 2017). The collection contains 122 texts, some texts contain more than one ambiguous pronoun. The passages are taken from books. They modified sentences to summarize context and backstory, to clarify or simplify, or to change gender or number of nouns and pronouns in order to introduce ambiguity. For challenge problems, character names were also changed. A single passage may give rise to multiple Pronoun Disambiguation Problems. Similar to the Winograd schemata, disambiguating pronouns in PDPs need a substantial amount of commonsense knowledge (see Example (9)).

- (9) *Sentence:* Always before, Larry had helped Dad with his work. But he could not help him now, for Dad said that his boss at the railroad company would not want anyone but him to work in the office.
Snippet: He could not help
Answer A: Larry
Answer B: Dad
Correct Answer: A

The texts are found in two splits: one is available on the web page as a snapshot of the whole collection (62 examples), the other is linked on the page (60 examples). The latter group is

⁷https://github.com/yandex-research/crosslingual_winograd

⁸<https://commonsensereasoning.org/disambiguation.html>



completely incorporated in the WNLI dataset of the GLUE benchmark (see Section 3.2), and one example from the former group is also included there.

Compared to Winograd schemata or other datasets mentioned in this paper PDPs are not formed as twin-sentences and they are much closer to real, natural texts and they are less artificial, however, they are still edited texts. A resource of unedited texts “from the wild” would be very advantageous, even so it is a well known fact, that processing, annotating unedited texts are always challenging.

4. HUNGARIAN WS AND WS-LIKE DATASETS

In this section we present the resources we built. Each of them was made on the basis of their English counterparts using machine translation and human validation. Our resources are gold standard quality, because the translation of every sentence were verified and corrected by linguistic experts with special regard to the phenomenon of pronominal anaphora.

However, before we turn to the newly created Hungarian resources, we will give a brief summary of the behavior of Hungarian pronouns, which we had to take into account during the translation process.

4.1. Hungarian pronouns

It is very important to note that although there are universal syntactic mechanisms behind the operation of pronominal anaphora and it can be traced back to general communication principles (e.g. maintaining coherence), the appearance and behavior of pronouns may differ from language to language. The Winograd schemata, the starting point of our work, contain English sentences and the basic literature on pronominal anaphora resolution mostly deal with the English language, but the preparation of the Hungarian dataset(s) requires a more thorough examination of the differences between Hungarian and English pronouns. When explaining these, we also emphasize the challenges and additional tasks they present in the case of the Hungarian language in resolving the anaphoric pronoun.

One of the important differences is that Hungarian is a pro-drop language, so in some cases pronouns can be left out from the sentence. In these cases, the person and number of the subject or object are calculable from the conjugation of the verb, or the person of the possessor from the conjugation of the possession. As in Example (10) there is no overt subject and object in the second sentence, yet it is clear who threw what.

- (10) A *gyerek* *játszott* *a* *labdával.* *Odadobta* *az*
 the kid play.PAST.SG3 the ball.Ins throw.PAST.DEF.SG3 the
apjának.
 father.POSS.SG3.DAT
 ‘The kid played with the ball. He tossed it to his father.’

Another important difference is that in contrast with English there is no grammatical gender in Hungarian. Thus, when searching for the antecedent of the third person singular pronoun, one can only rely on agreeing number and person as a surface feature. The semantic feature of animacy can help if we also have semantic analysis of the text, because the pronoun *ő* usually



refers to animate antecedent (11a), while the other pronoun *az* refers mostly to inanimate antecedent (11b), although *az* as a demonstrative pronoun can refer to animate antecedent as well (11c). The latter phenomenon appears typically in sentences where the subject of the second sentence is coreferent with any arguments of the first sentence except for the subject.

- (11) a. *Péter felhívta Marit, de ő nem vette fel.*
 Peter call.PAST.DEF.SG3 Mary.ACC but she not pick.PAST.DEF.SG3
fel.
 up
 ‘Peter called Mary, but she didn’t pick it up.’
- b. *Péter kikapcsolta a telefont, de az tovább csörgött.*
 Peter turn_off.PAST.DEF.SG3 the phone.ACC but it further
csörgött.
 ring.PAST.SG3
 ‘Peter turned off the phone, but it kept ringing.’
- c. *Péter felhívta Marit, de az nem vette fel.*
 Peter call.PAST.DEF.SG3 Mary.ACC but it not pick.PAST.DEF.SG3
fel.
 up
 ‘Peter called Mary, but she didn’t pick it up.’

4.2. HuWS: Hungarian winograd schemata

The original set of Winograd schemata was translated into Hungarian using machine translation. The output was validated by two linguists. Certain schemata were discarded, because they were not translatable preserving the features of the Winograd schemata. For instance, in the case of Example 12 we could not translate the phrases *breaking her silence*, *breaking her concentration* resulting a sentence pair that differ only in one word/phrase but preserving the possessive structure.

- (12) Lily spoke to Donna, breaking her [silence/concentration].

An other example of a sentence that was not translatable to Hungarian is in Example (4). The translation of ‘indiscreet’ and ‘noisy’ is the same in Hungarian: *indiszkrét*. Some schemata needed slight modifications for making them translatable.

It also had to be taken into account that pronouns behave differently across languages. In some cases (as in Example (13)) the target pronoun in the second clause is dropped in Hungarian (preserving the ambiguity of the target pronoun).

- (13) *A férfi nem tudta felemelni a fiát, mert olyan [gyenge / nehéz] volt.*
 the man not can.PAST.DEF.SG3 lift the son.POSS.SG3 because
 so weak heavy is.PAST.SG3
 ‘The man couldn’t lift his son because he was so [weak/heavy].’



No overt (personal or demonstrative) pronoun could appear in the second clause, when it would disambiguate the antecedent, as in Example (14).

- (14) *A férfi nem tudta felemelni a fiát, mert*
 the man not can.PAST.DEF.SG3 lift the son.POSS.SG3 because
*az olyan [*gyenge / nehéz] volt.*
 it so weak heavy is.PAST.SG3
 ‘The man couldn’t lift his son because he was so [*weak/heavy].’

The target pronoun is dropped in most of the schemata in Hungarian, however, there were some examples (e.g. Example (15)), where the structural ambiguity worked only with overt target pronoun. In these examples if the target pronoun were dropped, no structural ambiguity would appear, because it would refer back to the subject of the first clause. On the other hand, the overt target pronoun can refer back both to the subject and to the other nominal phrase as well.

- (15) *A tűzoltók a rendőrök [után / előtt] érkeztek*
 the fireman.PL the policeman.PL after before arrive.PAST.NDEF.PL3
ki, mert ők olyan messziről jöttek.
 out because they so far.DEL come.PAST.NDEF.PL3
 ‘The firemen arrived [after/before] the police because they were coming from so far away.’

We decided to change English proper names to Hungarian, except for the name of famous people. From the original dataset only 122 sentence pairs were translatable to Hungarian. The dataset is freely available.⁹

4.2.1. HuWNLI: Hungarian WNLI dataset. Following the practice of GLUE, the dataset has been transformed to an inference dataset, in order to provide suitable training data for neural models. Given a schema, as the one in Example (16), four sentence pairs were constructed from it by replacing the ambiguous pronoun with each possible referent (Example (17)). This way the task is formulized as (binary) sentence pair classification: the task is to predict if the second sentence (the one with the substituted pronoun) is entailed by the first sentence (thus the labels are *entailment* and *not-entailment*).

- (16) *A trófea nem fér bele a barna bőröndbe, mert túl [nagy/kicsi].*
 The trophy doesn’t fit into the brown suitcase because it is too [large/small].
- (17) a. *A trófea nem fér bele a barna bőröndbe, mert túl nagy. A trófea túl nagy.*
 The trophy doesn’t fit into the brown suitcase because it is too large.
 The trophy is too large.
 Label: entailment

⁹<https://github.com/nytud/HuWS>



- b. *A trófea nem fér bele a barna bőröndbe, mert túl nagy. A bőrönd túl nagy.*
The trophy doesn't fit into the brown suitcase because it is too large.
The suitcase is too large.
Label: not-entailment
- c. *A trófea nem fér bele a barna bőröndbe, mert túl kicsi. A trófea túl kicsi.*
The trophy doesn't fit into the brown suitcase because it is too small.
The trophy is too small.
Label: not-entailment
- d. *A trófea nem fér bele a barna bőröndbe, mert túl kicsi. A bőrönd túl kicsi.*
The trophy doesn't fit into the brown suitcase because it is too small.
The suitcase is too small.
Label: entailment

We inspected the three splits of the WNLI dataset and translated the sentence pairs that were not part of the original English schema collection. Those sentence pairs were mostly retrieved from the PDP dataset. One sentence pair of the training set was discarded as it was not translatable to Hungarian. We used these together with the above described Hungarian schemata (transformed into NLI format) as the Hungarian WNLI dataset. We followed the original splits of the WNLI dataset: 596 sentence pairs in the training set, 52 sentence pairs in the development set and 134 sentence pairs in the test set.¹⁰ We detected two erroneous labels in the training set (id 347 and id 464). We corrected them in our dataset. We also noticed that not every Winograd schema is included in the WNLI dataset (schema nr 22 and 29, for example). We inserted the Hungarian translations of those into the training set of HuWNLI. Human annotators applied labels to the test set of this dataset. The database is a part of the Hungarian Language Understanding Benchmark Kit (Ligeti-Nagy et al. 2022). The dataset is freely available.¹¹

4.3. PWS: parallel schemata including Hungarian

The original set of Winograd schemata is already translated to several languages now including Hungarian too. We attempted to align all translations available to one parallel resource, because on the basis of the parallel corpus interesting lessons can be drawn from pronouns as they behave differently in each language. In PWS, the translations (introduced in Section 3.3, including Portuguese, French, Mandarin, Japanese and Russian) and the Hungarian schemata are paired with the English original. In the case of Portuguese and Japanese two versions are involved: one preserving English proper names and one replacing them with names of the target language.

¹⁰The numbers may differ from the numbers of the WNLI dataset as some instances had to be discarded because they were not translatable to Hungarian, either during the translation of Winograd schemata, or during the process of translating the extra instances in the WNLI dataset.

¹¹<https://huggingface.co/datasets/NYTK/HuWNLI>



We tried to keep all additional information and metadata from the datasets. E.g. some translations provide the translator(s) of each schema, some information regarding to the translation, and the original dataset contains the source of the examples. Not every English schema was translatable to every language, therefore some fields of the file are left empty.

As it was mentioned in Section 4.2, during the translation process some sentences needed a more thorough rewording to keep the ambiguous pronouns and the two alternative antecedents. The same is certainly true of the other translations. It is important that if one uses this parallel corpus, be aware of the phenomenon.

The parallel dataset is available.¹²

4.4. HAPP: Hungarian ambiguous pronoun problems

All 1,882 sentences of the Definite Pronoun Resolution Dataset by Rahman & Ng (2012) were translated to Hungarian. In all sentence pairs of the Hungarian translation, either both target pronouns are dropped or none of them is. However, when translating the original sentence pairs, we have often been confronted with the fact that while one sentence contains an overt pronoun in the other, the two sentences cannot be unified in the sense that structural ambiguity is preserved and resolved only by commonsense reasoning.

There are conjunctions one can build a perfect schema for English with, but they behave differently in Hungarian. For instance, when translating a sentence pair like in Example (18), we cannot produce sentences where either both target pronouns are overt or both are dropped. This is especially true in sentence pairs where there is a conjunction *and* or *then* between the two clauses and the target pronoun is in subject role.

- (18) a. The bird ate the pie and it [died/was ruined].
 b. A madár belevett a pitébe, és [meghalt/*tönkrement].
 c. A madár belevett a pitébe, és az [*meghalt/tönkrement].

The solution is to split the sentence into two as in Example (19). By the way, the original English dataset of difficult pronouns contains 134 *and* conjunctions.

- (19) A madár belevett a pitébe. [Meghalt/Tönkrement].

If the target pronoun is not the subject of the second clause, the sentence is translatable without the above difficulty, as in Example (20).

- (20) a. The ball hit the window and Bill [repaired/caught] it.
 b. A labda eltalálta az ablakot, és Vili [megjavította/elkapta].

Other conjunction words like *mert* ('because'), *pedig* ('in turn'), *de* ('but'), *hogy* ('to'), *így* ('thus, so') are not problematic.

¹²<https://github.com/nytud/PWS>



In most cases the number of the target pronoun must agree with its antecedent. An exception is, however, cases where the antecedent is a collective name, because then the target pronoun can be plural, as in Example (21). Here the plural subject – expressed with a dropped pronoun – of the second clause can point back to either the plural object or the collective noun subject of the first clause.

- (21) a. The police arrested the rioters because they were [preventing/causing] trouble.
 b. A rendőrség letartóztatta a lázadókat, mert [megelőzték a bajt/bajt okoztak].

The original dataset was split into train and test sets following 30/70 ratio, we kept the same sets in the Hungarian version as well. The HAPP dataset is available.¹³

5. DISCUSSION

In this article, we have presented some resources that allow training and evaluating models that solve the problem of pronoun anaphora resolution. All of these new Hungarian datasets were created by translating their English antecedents. Our datasets are available under CC-BY-SA 4.0 license. We look forward to the results that can be achieved with the resources we have created.

In the next phase of the research, we plan adapting WinoGrande to Hungarian, and we would like to examine the possibility of building resources from living, real texts on the topic of ambiguous pronoun anaphora resolution.

REFERENCES

- Amsili, Pascal and Olga Semnck. 2017. A Google-proof collection of French Winograd schemas. Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017). 24–29. <https://doi.org/10.18653/v1/W17-1504>. <https://aclanthology.org/W17-1504>.
- Bernard, Timothée and Ting Han. 2020. Mandarinograd: A Chinese collection of Winograd schemas. Proceedings of the 12th Language Resources and Evaluation Conference. 21–26. <https://aclanthology.org/2020.lrec-1.3>.
- Copeland, B. Jack. 2000. The Turing test*. Minds and Machines 10. 519–539. <https://doi.org/10.1023/A:1011285919106>.
- Csendes, Dóra, János Csirik, Tibor Gyimóthy and András Kocsor. 2005. The Szeged Treebank. Proceedings of the 8th International Conference on Text, Speech and Dialogue. 123–131. https://doi.org/10.1007/11551874_16.
- Damassino, Nicola. 2020. The Questioning Turing Test. Minds and Machines 30(4). 563–587. <https://doi.org/10.1007/s11023-020-09551-6>.
- Davis, Ernest, Leora Morgenstern and Charles L. Ortiz. 2017. The first Winograd schema challenge at IJCAI-16. AI Magazine 38(3). 97–98. <https://doi.org/10.1609/aimag.v38i4.2734>. <https://ojs.aaai.org/index.php/aimagazine/article/view/2734>.

¹³<https://github.com/nytud/HAPP>



- Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) program – tasks, data, and evaluation. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). 837–840. <http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf>.
- Emelin, Denis and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 8517–8532. <https://doi.org/10.18653/v1/2021.emnlp-main.670>. <https://aclanthology.org/2021.emnlp-main.670>.
- Ghazal, Ahmad, Todor Ivanov, Pekka Kostamaa, Alain Crolotte, Ryan Voong, Mohammed Al-Kateb, Waleed Ghazal and Roberto V. Zicari. 2017. BigBench V2: The new and improved BigBench. 2017 IEEE 33rd International Conference on Data Engineering (ICDE). 1225–1236. <https://doi.org/10.1109/ICDE.2017.167>.
- Halliday, Michael A.K., Ruqaiya Hasan. 1976. Cohesion in English (A Longman Paperback). <https://doi.org/10.4324/9781315836010>. <https://books.google.hu/books?id=zMBZAAAAMAAJ>.
- Hirst, Graeme. 1981. Anaphora in natural language understanding: A survey. (Lecture Notes in Computer Science 119). Berlin & Heidelberg: Springer. <https://doi.org/10.1007/3-540-10858-0>.
- Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. Proceedings of the 37th International Conference on Machine Learning (PMLR) 119. 4411–4421. <https://doi.org/10.48550/arXiv.2003.11080>.
- Levesque, Hector J., Ernest Davis and Leora Morgenstern. 2012. The Winograd schema challenge. Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR'12. 552–561.
- Ligeti-Nagy, Noémi, Gergő Ferenczi, Enikő Héja, Kinga Jelencsik-Mátyus, László János Laki, Noémi Vadász, Zijian Győző Yang and Tamás Váradi. 2022. HuLU: Magyar nyelvű benchmark adatbázis kiépítése a neurális nyelvmodellek kiértékelése céljából [HuLU: Hungarian benchmark database to evaluate neural models]. XVIII. Magyar Számítógépes Nyelvészeti Konferencia. 431–446.
- de Melo, Gabriela S., Vinicius A. Imaizumi and Fabio G. Cozman. 2019. Winograd schemas in Portuguese. Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional. 787–798. <https://doi.org/10.5753/eniac.2019.9334>. <https://sol.sbc.org.br/index.php/eniac/article/view/9334>.
- Morgenstern, Leora, Ernest Davis and Charles L. Ortiz. 2016. Planning, executing, and evaluating the Winograd schema challenge. AI Magazine 37(1). 50–54. <https://doi.org/10.1609/aimag.v37i1.2639>. <https://ojs.aaai.org/index.php/aimagazine/article/view/2639>.
- Neufeld, Eric and Sonje Finnestad. 2020. In defense of the Turing test. AI & Society 35. 819–827. <https://doi.org/10.1007/s00146-020-00946-8>.
- Pradhan, Sameer, Mitch Marcus, Martha Palmer, Lance Ramshaw and Ralph Weischedel. 2007. OntoNotes: A unified relational semantic representation. Proceedings of the International Conference on Semantic Computing (ICSC 2007). 517–526. <https://doi.org/10.1109/ICOSC.2007.4338389>.
- Rahman, Altaf and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 777–789.
- Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula and Yejin Choi. 2019. WINOGRANDE: An adversarial Winograd schema challenge at scale. CoRR. abs/1907.10641. <https://doi.org/10.1145/3474381>. <http://arxiv.org/abs/1907.10641>.



- Shavrina, Tatiana, Alena Fenogenova, Anton Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. arXiv preprint. arXiv: 2010.15925. <https://doi.org/10.48550/arXiv.2010.15925>.
- Sun, Yu, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv. abs/2107.02137. <https://doi.org/10.48550/arXiv.2107.02137>.
- Tikhonov, Alexey and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 3534–3546. <https://doi.org/10.18653/v1/2021.findings-acl.310>
- Trichelair, Paul, Ali Emami, Adam Trischler, Kaheer Suleman and Jackie Chi Kit Cheung. 2019. How reasonable are common-sense reasoning tasks: A case-study on the winograd schema challenge and SWAG. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 3382–3387. <https://doi.org/10.18653/v1/D19-1335>. <https://aclanthology.org/D19-1335>.
- Turing, Alan. 1950. Computing machinery and intelligence. *Mind* 59(236). 433–460. <https://doi.org/10.1093/mind/LIX.236.433>.
- Vadász, N. 2020. KorKorpusz: kézzel annotált, többrétegű pilotkorpusz építése [KorKorpusz: manually annotated, multilayer pilot corpus]. In: XVI. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem TTK, Informatikai Intézet, pp. 141–154.
- Vincze, Veronika, Klára Hegedűs, Alex Sliz-Nagy and Richárd Farkas. 2018. SzegedKoref: A Hungarian coreference corpus. Proceedings of the 11th Language Resources and Evaluation Conference. 401–405.
- Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. 353–355. <https://doi.org/10.18653/v1/W18-5446>. <https://aclanthology.org/W18-5446>.
- Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy and Samuel R. Bowman. 2020. SuperGLUE: A stickier benchmark for general-purpose Language Understanding systems. Advances in Neural Information Processing Systems 32 (NeurIPS 2019). arXiv. 1905.00537v3. <https://doi.org/10.48550/arXiv.1905.00537>.
- Wang, Bochao, Hang Xu, Jiajin Zhang, Chen Chen, Xiaozhi Fang, Yixing Xu, Ning Kang, Lanqing Hong, Chenhan Jiang, Xinyue Cai, Jiawei Li, Fengwei Zhou, Yong Li, Zhicheng Liu, Xinghao Chen, Kai Han, Han Shu, Dehua Song, Yunhe Wang, Wei Zhang, Chunjing Xu, Zhenguo Li, Wenzhi Liu and Tong Zhang. 2020. VEGA: Towards an end-to-end configurable AutoML pipeline. arXiv. arXiv:2011.01507. <https://doi.org/10.48550/arXiv.2011.01507>.

Open Access. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited, a link to the CC License is provided, and changes – if any – are indicated. (SID_1)

