

HeGeL: A Novel Dataset for Geo-Location from Hebrew Text

Tzuf Paz-Argaman^{*,a}, Tal Bauman^{*,b}, Itai Mondshine^a,
Itzhak Omer^c, Sagi Dalyot^b, and Reut Tsarfaty^a

^aBar-Ilan University, Israel, ^bThe Technion, Israel, ^cTel Aviv University, Israel,
{tzuf.paz-argaman, mondshi1, reut.tsarfaty}@biu.ac.il,
talbauman@campus.technion.ac.il, omer@tauex.tau.ac.il,
dalyot@technion.ac.il

Abstract

The task of textual geolocation — retrieving the coordinates of a place based on a free-form language description — calls for not only grounding but also natural language understanding and geospatial reasoning. Even though there are quite a few datasets in English used for geolocation, they are currently based on open-source data (Wikipedia and Twitter), where the location of the described place is mostly implicit, such that the location retrieval resolution is limited. Furthermore, there are no datasets available for addressing the problem of textual geolocation in morphologically rich and resource-poor languages, such as Hebrew. In this paper, we present the Hebrew Geo-Location (HeGeL) corpus, designed to collect literal place descriptions and analyze lingual geospatial reasoning. We crowdsourced 5,649 literal Hebrew place descriptions of various place types in three cities in Israel. Qualitative and empirical analysis show that the data exhibits abundant use of geospatial reasoning and requires a novel environmental representation.¹

1 Introduction and Background

Textual Geolocation Identification, a crucial component of Geographic Information Retrieval (GIR), is the task of resolving the location, i.e., coordinates of a place, based on the reference to it in a text. It requires a combination of language and environmental knowledge. On top of the usual non-spatial linguistic challenges in Natural Language Understanding (NLU), such as named entity recognition (NER), anaphora resolution, bridging anaphora, etc., the textual geolocation task presents geospatial challenges that require multimodal processing and grounding (Ji et al., 2022; Fried et al., 2022; Misra et al., 2017; Qi et al., 2020; Paz-Argaman et al., 2020).

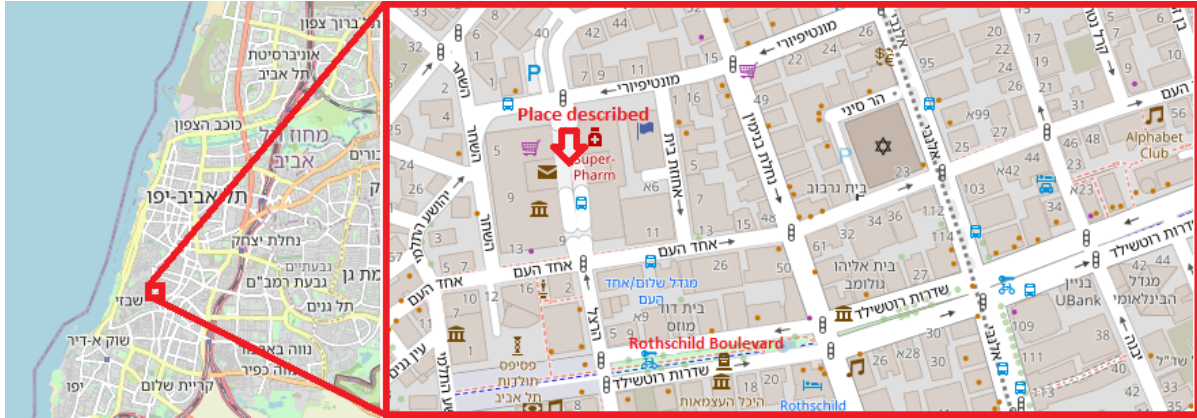
Proper names, such as ‘Rabin Square’, also known as *named entities* in Natural Language Processing (NLP), and as *rigid designators* in formal semantics (Kripke, 1972), can be easily grounded based on a Gazetteer or a simple map. However, geolocating linguistic terms that involve spatial expressions without the explicit mention of a proper name still present an open challenge. This interpretation challenge includes the understanding and resolution of (at least): (i) definite descriptions, such as ‘the school’ (ii) geospatial terms, such as cardinal directions; ‘east of’; and (iii) geospatial numerical reasoning; ‘two buildings away from the pharmacy’. To address these and other challenges, we need to both ground entity mentions to their corresponding physical entities in the environment, and to reason about geospatial relations expressed between entities — these two processes being closely intertwined.

To do so, we need a corpus for the geolocation task that maps rich geospatial place descriptions to their corresponding location coordinates. However, current corpora for geolocation are based on naturally-occurring open-source resources, such as Wikipedia articles (Eisenstein et al., 2010; Wing and Baldrige, 2011; Han et al., 2012; Wing and Baldrige, 2014; Wallgrün et al., 2018), which are not spatially oriented, i.e., the description of locations is implicit or absent in the corresponding text. Subsequently, the accuracy of retrieval is fairly low (around 100 km).

Furthermore, all geolocation datasets previously studied in NLP are in English, with a dearth of corpora for low-resource languages, in particular, for morphologically rich languages, such as Hebrew. To understand the geolocation challenges and build models that do various spatial reasoning tasks, English cannot be our sole focus (Baldrige et al., 2018). Hebrew, a Semitic morphologically rich language is notoriously difficult to parse (Tsarfaty et al., 2020, 2019). Moreover, resources that are

^{*}Equal contribution.

¹For data and code see <https://github.com/OnlpLab/HeGeL>



Place Description: The place is located near the Rothschild complex – at the end of Rothschild Street, as you go towards the sea, take a right for about three streets and then you will see the tower high above you.

Figure 1: A description example from HeGeL translated from Hebrew.

available for Hebrew NLP research focus on traditional tasks, such as Part-of-speech (POS) tagging, syntactic parsing, etc; and lack corpora for understanding and reasoning in real-world situations.

In this work we present HeGeL, a novel dataset for **Hebrew Geo-Location**, the first ever Hebrew NLU benchmark involving both grounding and geospatial reasoning. To create HeGeL, we crowdsourced 5,649 geospatially-oriented Hebrew place descriptions of various place types from three cities in Israel. We designed our task based on a realistic scenario of human place description, relying on people’s memory of the world, rather than, e.g., using a map (Anderson et al., 1991; Paz-Argaman and Tsarfaty, 2019). Crucially, relying on environmental cognition results in various levels of geospatial knowledge (Siegel and White, 1975) that are manifested in the descriptions and the geospatial reasoning that is required to resolve their location (Hayward and Tarr, 1995). To avoid the much simpler task of grounding proper named entities, we explicitly restricted the use of proper names in the description of the place and adjacent landmarks.

Unlike the text-based navigation task (MacMahon et al., 2006; Chen et al., 2019; Ku et al., 2020; De Vries et al., 2018; Thomason et al., 2020), which requires representing an agent’s current perspective, reflecting its route knowledge, we show that the HeGeL task requires a full-environment representation, thus, capturing complex geospatial relations among multiple physical entities. Through a thorough linguistic and empirical analysis, we demonstrate the characteristics and challenges associated with Hebrew place descriptions,

showing that HeGeL serves both as a challenging NLU benchmark and as a corpus for geospatial cognition research.

2 The HeGeL Task and Dataset

This work addresses the task of geolocating places on a map based on natural language (NL) geospatial descriptions that are given in a colloquial language and based on participants’ memory of the environment (i.e., cognitive map). The input to the HeGeL task is as follows: (i) an NL place description of the whereabouts of the place, and (ii) a map with rich details of the environment (e.g., physical entities names, geospatial relations, and attributes). The output is a pair of coordinates (x,y) specifying the physical location of the place described in the text. Figure 1 shows an example of a place description from HeGeL translated from Hebrew.

To simplify the crowdsourcing task and encourage participants’ engagement, we frame the data crowdsourcing process as the well-known game, the *treasure hunt* task (Kniestedt et al., 2022), in which the *instructor-participant* is required to describe in writing the location of the treasure, a known place in the city, to a different *follower-participant* who then needs to locate it on a map. Thus, the online assignment is divided into two tasks: the instructor’s *writing of place descriptions* and the follower’s *validation*. To avoid preconceived notions as to the ‘correct’ way to describe a place, we first presented the participants with the task of writing a place description, and once completed, the validation task was given.²

²Appendix A includes additional data collection details.

We hereby provide the details of the two UI tasks:

(i) Task 1. Writing a place description In this task we requested participants to describe in a free-form text the location of a place known to them, to a third party who might not be familiar with the whereabouts of that place. To collect place descriptions based solely on people’s memory, we did not visualize the area of the place, e.g., on a map. Instead, we ensured that the participants are well familiarized with the place by asking them to state how familiar they are with the place on a scale of 1-5. If this score was 1 or 2, we presented the participant with a different place to describe. To ensure diverse human-generated textual descriptions, places were chosen based on their type, position/location in the city (places were spread across the city), geometry, size, and context. To avoid the use of proper names, we developed a rule-based methodology to make sure that the explicit name of the goal (place) or of the nearby landmarks (< 100 meters) will not appear explicitly in the description. The original description was saved, and the participants were asked to input another description without the above names.

(i) Task 2. Place description validation To verify that a person who reads the text description will understand where the treasure is hidden, i.e., geolocate the place, we developed a map-based retrieval task. The participant in the follower role was asked to read the crowdsourced textual description and mark its location on the map, i.e., where the treasure is hidden. For marking the location, we implemented an interactive online map based on [OpenStreetMap \(OSM\)](#),³ which allows the participants to move and zoom-in to precisely pin the described place on the map. The map supports the cognitive process needed to ground mentioned entities to physical entities, reason about the geospatial relations, and locate the described place. To familiarize participants with the interactive map tool and task, they had to first pass a simple map marking test, and only then they could start task 2 of reading place descriptions (given by other participants), marking place locations on the map, and rate the clarity of the textual description on a scale of 1-5.

Appendix B presents a display of the online assignment’s UI translated from Hebrew to English.

³OSM is a free, editable, map of the whole world, that was built by volunteers, with millions of users constantly adding informative tags to the map.

Target Selection and Retrieval Errors The treasure-hunt task we devised included 167 places in the three largest cities in Israel: Tel Aviv, Haifa, and Jerusalem. These three cities are differently shaped, and show different physical, morphological and topographic features, which potentially affect the legibility and imageability of urban components, and therefore also on place descriptions. These differences can be expressed in the use of various physical features and prepositions, e.g., frequent use of the physical object ‘landmark’ and the prepositions ‘above’ or ‘below’ in hilly terrains that characterize Haifa and Jerusalem.

To assess the quality and interpretability of the place descriptions, we calculate the shortest Euclidean distance between the coordinates of the goal’s (physical element) shape (polygon, line or point), and the location marked by the ‘follower’ on the map (task 2); we term this distance as *retrieval error*. To determine the agreement rate among human participants, each textual place description is validated by at least two participants. To ensure that we work with descriptions that can be geolocated, we set a hard distance threshold of 300 meters, based on analysis of the descriptions’ clarity score that we had conducted on a prior (held-out) development corpus we collected for the task.

3 Data Statistics and Analysis

The resulting HeGeL dataset contains 5,649 validated descriptions paired with their coordinates on a map. The locations are divided among three cities: 2,142 in Tel Aviv, 1,442 in Haifa, and 2,065 in Jerusalem. 1,833 participants completed the writing task, inserting in total 10,946 place descriptions, and 2,050 participants completed 12,655 validation tasks. The dataset is balanced, with about 33 descriptions per place.

Figure 2 shows a Venn diagram representing the relation of the three sets of city-based vocabularies (formed from unique lemmas produced by [More et al. \(2019\)](#) lemmatization tool). The intersection of the three cities contains only 15.07% of the entire vocabulary (the union of the three cities’ vocabularies). The shared language is not focused on city-specific terms, such as ‘Knesset’. Instead, it includes rich spatial terms, such as ‘between’, modified prepositions such as ‘next to’, and non-definite entities, such as ‘street’. From the Venn diagram we also conclude that almost half of the lemmas of the three vocabularies, corresponding

Phenomenon		c	μ	Example from HeGeL (translated into English)
Type of elements in a city (Lynch, 1960)	Edge	36%	0.6	“when reaching Yafo, one should go toward the sea...”
	Node	40%	0.44	“... a few minutes walk from the HaShaon square...”
	Landmark	60%	1.08	“... near Levinski market”
	District	36%	0.4	“South part of the city next to...”
	Path	68%	0.76	“On Carlebach street...”
Spatial knowledge (Siegel and White, 1975)	Landmarks	32%	n/a	“Next to the sea in Tel Aviv-Yafo”
	Route	20%	n/a	“Passing Azrieli on Menachem Begin and then turn right...”
	Survey	48%	n/a	“South part of the city near Levinski market”
Reference to unique entity		100%	2.32	“... in the middle of Dizengoff street”
Cardinal direction		44%	0.76	“South of Sharona...”
Coreference		16%	0.16	“... continue a bit west and it...”

Table 1: Linguistic qualitative analysis of 25 randomly sampled descriptions in HeGeL. c is the percentage of descriptions containing at least one example of the phenomenon, and μ is the mean number of times the phenomenon appears in each description.

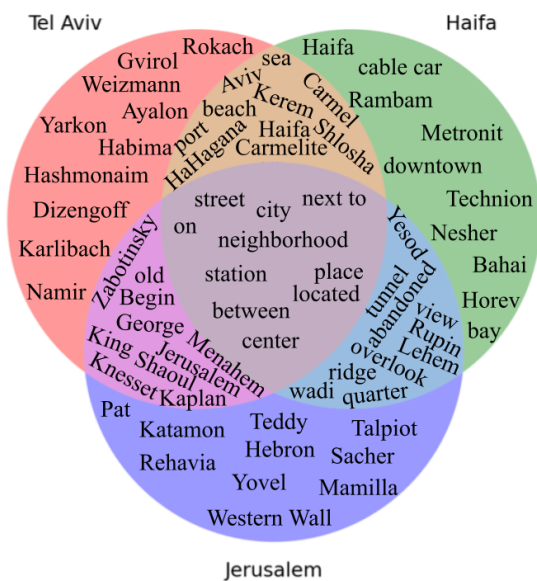


Figure 2: A Venn diagram showing the top 10 words (translated from Hebrew) used in each city-set relation.

to the three cities, contain city-specific lemmas: 48.6%, 40.65%, and 49.3% for Tel Aviv, Haifa, and Jerusalem, respectively. As such, HeGeL enables a city-split setup, training on one city and testing on a different unseen city, where city-reserved named entities present an out-of-vocabulary (OOV) challenge for models trained on another city.

Table 1 shows an analysis of the linguistic phenomena manifested in the HeGeL dataset, demonstrating the spatial knowledge and reasoning skills required for solving the HeGeL task. We analyzed the frequency of the five types of elements in a city defined by Lynch (1960), along with the three types of spatial knowledge defined in Siegel and White (1975), and other spatial properties. The frequent use of cardinal directions, as well as the use of sur-

Feature	Avg. per description	Unique in corpus
Number of lemmas	12.93	6,663
Number of tokens	11.50	9,207
Number of named entities	0.55	3,490
Number of prepositions	2.39	14,256
Number of verbs	0.53	3,152

Table 2: Quantitative analysis of HeGeL.

Feature		p-value	FDR corrected p-value
Task 1:	Number of Words	0.8306	0.8306
	Number of named entities	0.0000	0.0000
	Place description (linguistic features)	Number of prepositions	0.0145
Task 2:	Number of verbs	0.3400	0.4080
	Retrieval error	0.0000	0.0000
Human verification	Clearness score	0.0000	0.0000

Table 3: Correlations between place types and linguistic and features.

vey knowledge, suggests that any NLP model built to deal with the HeGeL task should not only represent a local view of the goal, or possible routes, but also take into consideration the full region, and mimic people’s map-like view of the environment. Therefore, unlike navigation tasks where only the agent’s current perspective is represented in the model, this task requires full representation of the environment.

We further perform a quantitative analysis of word tokens and lemmas that appear in HeGeL, depicted in Table 2. Overall, the HeGeL dataset contains a large vocabulary of 9,207 unique tokens and 6,663 unique lemmas. There are mentions of physical entities, but as we limited the mentions of named-entities of the described place and landmarks adjacent to it; these are relatively rare, and are mostly references to prominent city landmarks. Also, as most place descriptions are not route-based descriptions, there are only few verbs used in the descriptions. Prepositions, on the other hand, are abundant.

In Table 3, using a one-way analysis of variance (ANOVA) test, we found a significantly ($p < 0.05$) different distribution between place type descriptions and the following features: number of named entities, number of verbs, human verification retrieval error, and clarity score.

4 Experiments

We create a zero-shot (ZS) city-based split, such that we train on one city and test on another. The train, development, and test sets correspond to the descriptions collected in Tel Aviv, Haifa, and Jerusalem, respectively. We evaluate different baseline models for the geolocation task on the HeGeL dataset. We use three evaluation metrics based on retrieval error: mean, median, and task completion (TC) accuracy – the percentage of place descriptions located within the 300 meters threshold. We provide three baselines for the HeGeL task.

We first assess a brute-force NER approach; i.e., we test whether recognizing named entities in the text and retrieving their corresponding coordinates is sufficient for solving the HeGeL task of geolocation. To this end, we used Google Maps API and produced two baseline models: (i) Google Maps API Query — we queried the API with the full raw text descriptions as input, with no preprocessing; and (ii) Oracle NER — we queried all 1-5 n-grams against Google Maps API and retrieved the closest geolocation to the goal.

In our second approach, we employ a dual-encoder model. One encoder encodes the text using a Hebrew Monolingual pre-trained encoder, AlephBERT (Seker et al., 2022), which produces a 768-dimension vector representation of the text. The other encoder processes the environment, which is represented as a graph based on OSM data. Each point of interest in the graph is connected to an S2Cell⁴, which contains its geometry and is based on S2-geometry. These S2Cells are encoded using a random-walk algorithm to produce a 64-dimensional vector for each cell. These vectors are then passed through a linear layer to produce 768-dimensional vectors. We calculate the cosine similarity score between the text and environment vectors and use it to align the respective representations via maximization of the cosine similarity score with a cross-entropy loss over the scores.

⁴S2Cells are based on S2-geometry (<https://s2geometry.io/>), a hierarchical discretization of the Earth’s surface (Hilbert, 1935).

Split	Model	Mean	Median	TC
ZS	Google Maps API Query	2,811	849	27.66
	Oracle NER*	2,373	496	37.79
	HUMAN	553	151	70.81**
ZS		2,727(1684)	2,612(1930)	2.37(1.5)
FS 20%	Dual-encoder	1717(35)	1583(49)	3.43(0.09)
FS 80%		983(23)	632(13)	15.7(0.38)

Table 4: Baseline results over the zero-shot (ZS) city-split, and few-shot (FS) split of different sizes: 20% and 80% of the samples in the test-region. For the Dual-encoder we report the mean over three random initialization and the standard-deviation (std) is in brackets. *Oracle NER is a skyline model based on a NER approach. **The human agreement rate.

Performing an ANOVA test, we found a significantly ($p < 0.05$) different distribution between place type descriptions and the retrieval error of the Oracle NER. The human retrieval error of the Path and Node place types were the lowest in both human verification and Oracle NER. This suggests that both of these place types are easier for humans to geolocate.

The results in Table 4 show that our task is not solvable with adequate resolution by the Google Maps API. The human performance provides an upper bound for the HeGeL task performance, while the simple Google Maps API Query provides a lower bound. The Google API model’s low performance suggests that NER and the Gazetteer-based methods in and of themselves are insufficient to handle the HeGeL task successfully, and that geospatial reasoning is necessary. The Dual-encoder’s low performance on the ZS split suggests that OOV is a major challenge. The few-shot (FS) split shows an improvement of the model after fine-tuning on additional samples from the test-region (FS 20% and 80%). This suggests that a possible solution for the city-split setup might be data-augmentation via generating grounded descriptions for the tested region – an approach we reserve for future research.

5 Conclusion

The contribution of this paper is threefold. First, we present the first geolocation benchmark with Hebrew place descriptions. Second, to the best of our knowledge, this is the only *crowdsourced* geolocation dataset, thus, eliciting explicit geospatial descriptions, allowing for better retrieval resolution. Finally, our analysis shows that the dataset presents complex spatial reasoning challenges which require novel environmental model representation.

Limitations

While we aim for our HeGeL crowdsourcing methodology to be applicable to other languages, and in particular low-resource languages, the UI design and our analyses require knowledge of the intended language, as well as familiarity with the regions where it is spoken. Moreover, as our methodology relies on people’s familiarity with the places, it limits the cities chosen for the task and the participants that could take part, restricting the demographics of the participants accordingly. In addition, relying on people’s memory of the environment causes many of the descriptions to be too vague for humans to geolocate, thus, many of the descriptions were disqualified during the validation process as they could not have been resolved. The relatively low percentage of place descriptions that were successfully validated, raises the costs of collecting such a dataset.

Acknowledgements

This research is funded by a grant from the European Research Council, ERC-StG grant number 677352, and a grant by the Israeli Ministry of Science and Technology (MOST), grant number 3-17992, for which we are grateful.

References

- Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hrc map task corpus. *Language and speech*, 34(4):351–366.
- Jason Baldrige, Tania Bedrax-Weiss, Daphne Luong, Sridhar Narayanan, Bo Pang, Fernando Pereira, Radu Soricut, Michael Tseng, and Yuan Zhang. 2018. Points, paths, and playscapes: Large-scale spatial language understanding tasks set in the real world. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 46–52.
- Howard Chen, Alane Suhr, Dipendra Misra, Noah Snaveley, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.
- Harm De Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*.
- Jacob Eisenstein, Brendan O’Connor, Noah A Smith, and Eric Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1277–1287.
- Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2022. Pragmatics in grounded language learning: Phenomena, tasks, and modeling approaches. *arXiv preprint arXiv:2211.08371*.
- Bo Han, Paul Cook, and Timothy Baldwin. 2012. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of COLING 2012*, pages 1045–1062.
- William G Hayward and Michael J Tarr. 1995. Spatial language and spatial representation. *Cognition*, 55(1):39–84.
- David Hilbert. 1935. Über die stetige abbildung einer linie auf ein flächenstück. In *Dritter Band: Analysis-Grundlagen der Mathematik- Physik Verschiedenes*, pages 1–2. Springer.
- Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert D Hawkins, and Yoav Artzi. 2022. Abstract visual reasoning with tangram shapes. *arXiv preprint arXiv:2211.16492*.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Isabelle Kniestedt, Iulia Lefter, Stephan Lukosch, and Frances M Brazier. 2022. Re-framing engagement for applied games: A conceptual framework. *Entertainment Computing*, 41:100475.
- Saul A Kripke. 1972. Naming and necessity. In *Semantics of natural language*, pages 253–355. Springer.
- Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldrige. 2020. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*.
- Kevin Lynch. 1960. The image of the environment. *The image of the city*, 11:1–13.
- Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. *Def*, 2(6):4.
- Dipendra Misra, John Langford, and Yoav Artzi. 2017. Mapping instructions and visual observations to actions with reinforcement learning. *arXiv preprint arXiv:1704.08795*.
- Amir More, Amit Seker, Victoria Basmova, and Reut Tsarfaty. 2019. Joint transition-based models for morpho-syntactic parsing: Parsing strategies for mrls and a case study from modern hebrew. *Transactions of the Association for Computational Linguistics*, 7:33–48.

- Tzuf Paz-Argaman, Yuval Atzmon, Gal Chechik, and Reut Tsarfaty. 2020. Zest: Zero-shot learning from text descriptions using textual similarity and visual summarization. *arXiv preprint arXiv:2010.03276*.
- Tzuf Paz-Argaman and Reut Tsarfaty. 2019. Run through the streets: A new dataset and baseline models for realistic urban navigation. *arXiv preprint arXiv:1909.08970*.
- Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991.
- Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. Alephbert: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–56.
- Alexander W Siegel and Sheldon H White. 1975. The development of spatial representations of large-scale environments. *Advances in child development and behavior*, 10:9–55.
- Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR.
- Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From spmrl to nmrl: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (mrls)? *arXiv preprint arXiv:2005.01330*.
- Reut Tsarfaty, Amit Seker, Shoval Sadde, and Stav Klein. 2019. What’s wrong with hebrew nlp? and how to make it right. *arXiv preprint arXiv:1908.05453*.
- Jan Oliver Wallgrün, Morteza Karimzadeh, Alan M MacEachren, and Scott Pezanowski. 2018. Geocorpora: building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, 32(1):1–29.
- Benjamin Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 955–964.
- Benjamin Wing and Jason Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 336–348.

A Data Collection Details

We used the services of an Israeli surveying company to distribute the assignment to native Hebrew-speakers participants in Israel only. The survey company was charged with distributing the assignments to a balanced set of participants in terms of their demographic and geographic characteristics (e.g., an equal number of males and females). All participants were given full payment, non-respective of whether they correctly completed the task.

The first page the participants viewed contains a disclosure about the assignments being part of academic research and the purpose of the assignments. The assignment protocol was approved by a behavioral review board. This approval was also presented to the participants on the initial screen. Also, the participants were required to read an informed consent form and sign an agreement box.

B Participant Interface

The tasks are performed via an online assignment application, depicted in Figures 3-5.

C Experimental Setup Details

The cross-entropy loss function was optimized with Adam optimizer (Kingma and Ba, 2015). The hyperparameter tuning is based on the average results run with three different seeds. The Learning rate was searched in [1e-5, 1e-4, 1e-3] and a 1e-5 was chosen. The S2cell level was searched in [13, 15, 17] and 13 was chosen. Number-of-epochs for early stopping was based on their average learning curve.

Play "treasure hunt"!

Your task is to describe to a friend (who has a map) where in the city the treasure is hidden. As your friend does not know well the city (Tel-Aviv), you must provide a detailed geographic description where the place is located, so the friend can find the place once he arrives to the area

Example of a good description: **V** - Located in North Herzl Street, East of Ayalon River, in the Emek hospital

Example of a bad description: **X** - In a big building with stores

Rules of the game:

- Do not mention the exact address.
- Do not mention the name of the place.
- Describe where it is, not what it is.
- Descriptions must contain at least 6 words.

Attention! Descriptions will be checked – descriptions not done by the rules will not be eligible for points

Figure 3: Participant Interface translated from Hebrew: instructions for the writing task.

Know the place?

The Grand Synagogue ★

How well do you know the place?

▼
Select

Select

Do not know at all - 0

1

2

3

4

know very well - 5

Describe where the place is

SEND DESCRIPTION


Not sure? Find another place

Figure 4: Participant Interface translated from Hebrew: the writing task (task 1).

In the following page you will be presented with 20 location descriptions in Tel Aviv

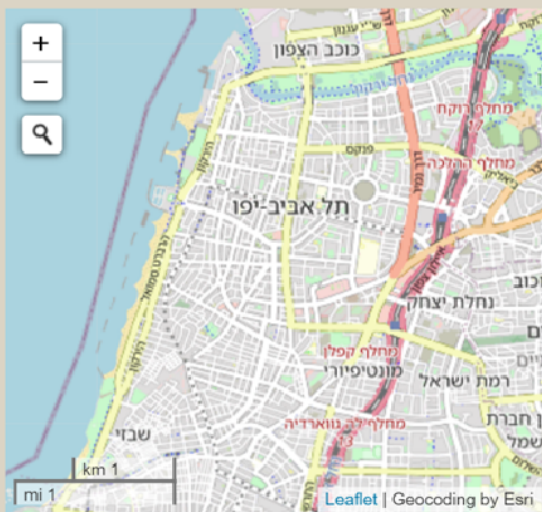
Below each description there is an interactive map that you can pan/move, zoom in/out - .etc

If you recognize the first place described, mark it on the map in maximum zoom, and write how clear the description was. Otherwise enter 0

To find the places mentioned in the description you can use the search button 

Here is an example - mark the following location

Your destination is the Azrieli Center. It is East of Menachem Begin Rd., and West of Ayalon South Rd., next to Kaplan and Hashalom interchanges



▼ Select

Let's go

Figure 5: Participant Interface translated from Hebrew: the validation task (task 2).

ACL 2023 Responsible NLP Checklist

A For every submission:

- A1. Did you describe the limitations of your work?
limitation section
- A2. Did you discuss any potential risks of your work?
The data doesn't contain private or sensitive information
- A3. Do the abstract and introduction summarize the paper's main claims?
1
- A4. Have you used AI writing assistants when working on this paper?
Left blank.

B Did you use or create scientific artifacts?

Left blank.

- B1. Did you cite the creators of artifacts you used?
No response.
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C Did you run computational experiments?

4

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
The model is a standard encoder-decoder run for around 10 minutes with early stopping on 1 GPU.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
in the appendix
- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4
- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3
- D** **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
2
- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
in the appendix
- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
in the appendix
- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
in the appendix
- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
in the appendix
- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
in the appendix