# Machine Translation between Hebrew and Arabic: Needs, Challenges and Preliminary Solutions


| Reshef Shilon | Nizar Habash | Alon Lavie | Shuly Wintner |
|---|---|---|---|
| Dept. of Linguistics | CCLS | LTI | Dept. of Computer Science |
| Tel Aviv U. | Columbia U. | Carnegie Mellon U. | U. of Haifa |
| Tel Aviv, Israel | New York, NY | Pittsburgh, PA | Haifa, Israel |


Modern Hebrew and Modern Standard Arabic, both Semitic languages, share many orthographic, lexical, morphological, syntactic and semantic similarities, but they are still not mutually comprehensible. Most native Hebrew speakers in Israel do not speak Arabic, and the vast majority of Arabs (outside Israel) do not speak Hebrew. Machine translation (MT) between these two language has the potential to bridge over political and cultural differences and bring the disputing peoples in the Middle East somewhat closer together by better understanding each other's societies.

The dominant paradigm in contemporary MT (Brown et al., 1990) relies on large-scale parallel corpora from which correspondences between the two languages can be extracted. However, such abundant parallel corpora currently exist only for few language pairs; and low- and medium-density languages (Varga et al., 2005) require alternative approaches. Specifically, no parallel corpora exist for Hebrew–Arabic.[1]

As an alternative to the pure statistical approach, we are currently developing a Hebrew-to-Arabic MT system, using the Stat-XFER framework (Lavie, 2008), which is particularly suited for low-resource language pairs. We discuss some linguistic properties of the two languages. We describes the implications on MT of the similarities and, in particular, differences between the two languages. We then discuss possible solutions to these challenges, advocating a linguistically-aware, transfer-based approach. Finally, we describe the system we are in the process of developing and reports some preliminary results.

## References


P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

Alon Lavie. Stat-XFER: A general search-based syntax-driven framework for machine translation. In Alexander F. Gelbukh, editor, *CICLing*, volume 4919 of *Lecture Notes in Computer Science*, pages 362–375. Springer, 2008. ISBN 978-3-540-78134-9.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. In *Proceedings of RANLP'2005*, pages 590–596, 2005.


[1] Several web sites have *comparable* contents, e.g., Wikipedia or the Israeli daily YNet (`http://www.ynet.co.il`); A small set of translated political essays is available from Gush Shalom (`http://www.gush-shalom.org/`) and Zavit Akheret (`http://zavita.co.il/`); the Bible is not available in Modern Hebrew.