

# MOSES: BRINGING MACHINE TRANSLATION TO THE MASSES



Machine translation (MT) is a research field that investigates the use of a computer to translate from one natural language to another. It has obvious practical benefits in enabling people to communicate with others who do not share a common language.

Most modern MT research focuses on the use of machine learning and statistical techniques to create translation systems – the dominant approach is known as ‘statistical machine translation’ (SMT). This is the approach adopted by the online translation systems offered by Google and Microsoft.

SMT requires large numbers of reliable language-specific translation examples. The SMT system then analyses the translations and matches the phrases and words that correspond in each language and uses these matches to guide future translations.

The system can continue to learn as it is fed more translation examples and can be tailored to specific fields.



## IMPACT: FROM POLITICS TO CHAT

Moses provides increased productivity and lower prices for companies that use it, and has also helped to open up new markets for machine translation.

Services such as the European Commission’s Europe Media Monitor (EMM) translate more than 100,000 articles a day in 50 languages for dissemination within the Commission. The EMM has used Moses since 2009.

Electronic discovery (e-discovery) is the digital forensic analysis of vast amounts of information during litigation and commercial transactions, such as company takeovers, in order to find relevant information. The global economy has increased the need for high-speed bulk translation of foreign-language documents and emails during this process.

Technology providers such as Simple Shift use Moses as the underlying technology to build translation systems for this market.

Translation of real-time interactive chat and near-real-time translation of user reviews, public forums, and bulletin boards have been demonstrated. Multilingual interaction cannot afford the luxury of expert human post-revision but they must be of sufficient quality.

In contrast to general-purpose translation services such as Google and Bing, systems built on Moses can be trained on user-specific and domain-specific data, resulting in better translation quality.



In 2005, the Edinburgh MT group developed the Moses toolkit using a statistical approach. This toolkit has been one of the main drivers in making MT more accessible to small and medium-sized companies.

The development of Moses has led to a significant increase in the understanding and use of MT in the translation industry. The free, open-source licence has allowed many organisations to access the latest developments in MT research that had once been the preserve of governments and large IT companies.

Unlike some commercial software, Moses can be installed on a user's own server and so can translate commercially and politically sensitive material without the additional security risks of sending the material outside the company.

The quality of the Moses translation depends on the type and amount of material used to train the system. This initial setup makes the toolkit highly customisable. For instance, a legal firm can train the system using Dutch-English legal translation examples, resulting in much better translations than a system trained for general purpose.

Moses is one of the most widely adopted MT systems in the translation industry. Its maturity and quality, as well as its liberal open-source licence, means that it is often preferred over proprietary systems.

The toolkit is continuously being developed to improve its efficiency and usability, and to incorporate advances in MT research. University of Edinburgh researchers are at the forefront of the development of the toolkit.



The Europe Media Monitor uses Moses.

## RELEVANT PUBLICATIONS FROM THE MT GROUP

Moses: Open Source Toolkit for Statistical Machine Translation, Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A and Herbst E, Association for Computer Linguistics 2007.

[www.aclweb.org/anthology/P07-2045](http://www.aclweb.org/anthology/P07-2045)

Factored Translation Models, Koehn P and Hoang H, EMNLP 2007.

[www.aclweb.org/anthology/D07-1091](http://www.aclweb.org/anthology/D07-1091)

Agreement Constraints for Statistical Machine Translation into German, Williams P and Koehn P, Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT), 2011.

[www.aclweb.org/anthology/W11-2126](http://www.aclweb.org/anthology/W11-2126)

Enriching Morphologically Poor Languages for Statistical Machine Translation, Avramidis E and Koehn P, ACL 2008.

[www.aclweb.org/anthology/P/P08/P08-1087](http://www.aclweb.org/anthology/P/P08/P08-1087)

Improving Mid-Range Reordering using Templates of Factors, Hoang H and Koehn P, EACL 2009.

[www.aclweb.org/anthology/E09-1043](http://www.aclweb.org/anthology/E09-1043)

Monte Carlo Inference and Maximization for Phrase-based Translation, Arun A, Dyer C, Haddow B, Blunsom P, Lopez A and Koehn P, Conference on Computational Natural Language Learning, 2009.

[www.aclweb.org/anthology/W09-1114](http://www.aclweb.org/anthology/W09-1114)

## CONTACT

Professor Philipp Koehn  
Chair of Machine Translation  
School of Informatics  
The University of Edinburgh

[pkoehn@inf.ed.ac.uk](mailto:pkoehn@inf.ed.ac.uk)

**If you require this document in an alternative format, such as large print, please contact:**

Communications and Marketing

Tel: +44 (0)131 650 2252

Email: [communications.office@ed.ac.uk](mailto:communications.office@ed.ac.uk)



THE UNIVERSITY of EDINBURGH