# Le système STAM

Mehdi Embarek[1]
(1) MK SOFT, 11 rue des fossés St Marcel, 75005 Paris
embarekm@gmail.com

**Résumé.** Le projet STAM aborde la problématique de la transcription automatique du langage texto (SMS) et plus particulièrement la traduction des messages écrits en arabe dialectal. L'objectif du système STAM est de traduire automatiquement des textes rédigés en langage SMS dans un dialecte parlé dans le monde arabe (langue source) en un texte facilement interprétable, compréhensible et en bon français (langue cible).

**Abstract.** The STAM project addresses the problem of automatic transcription of SMS language and especially the translation of messages written in Arabic dialect. The objective of STAM system is to automatically translate texts written in SMS language in a dialect spoken in the Arab World (source language) into a French text (target language), interpretable and understandable.

**Mots-clés :** Dialecte, SMS, Transcription, STAM.
**Keywords:** Dialect, SMS, Transcription, STAM.

L'évolution des technologies de communication a permis de développer différentes formes de langage. En plus du langage naturel utilisé dans tous les documents écrits, un autre langage, que l'on retrouve dans des documents ou des textes informels, a vu le jour. Il s'agit du langage SMS[1] (ou langage texto) qu'il faudra dorénavant considérer comme un nouveau langage à part entière. Aussi, depuis l'expansion des réseaux sociaux, des blogs et des forums de discussions, les internautes utilisent de plus en plus ce langage pour exprimer leur avis concernant une actualité ou commenter un évènement. Et pour certaines communautés (maghrébine par exemple), on emploie même dans les messages des termes issus de leur dialecte local. Ce qui rend la compréhension du texte presque impossible pour les personnes ne parlant pas le dialecte employé. Ces textes non structurés peuvent être considérés comme des sources d'informations et il serait intéressant de pouvoir les exploiter et les analyser afin d'en extraire le contenu informationnel en se basant sur des outils et techniques adaptés. De nombreux travaux ont été menés dans ce sens afin d'étudier plus particulièrement les spécificités des dialectes (Guella, 2011) (Vanhove, 1999) ou encore d'élaborer des terminologies (dictionnaires) propre à chaque pays (www.speakmoroccan.com pour le Maroc, www.arabetunisien.com pour la Tunisie). Cependant, peu de travaux se sont penchés sur le développement ou la proposition d'un système de traduction automatique.

Le projet STAM (Système de Transcription AutoMatique) (http://www.stam-dz.com) est un projet de recherche industrielle soutenu conjointement par la société Med Point Dz et la société MK Soft. Le projet aborde la problématique de la transcription (traduction) automatique du langage texto (SMS) et plus particulièrement la traduction des messages écrits en arabe dialectal. L'objectif du système STAM est de traduire automatiquement des textes rédigés en langage SMS dans un dialecte parlé dans le monde arabe (langue source) en un texte facilement interprétable, compréhensible et en bon français (langue cible).

Dans notre développement, nous nous sommes particulièrement intéressés au dialecte algérien, c'est-à-dire à tous les dialectes parlés en Algérie (l'algérois, l'oranais, le constantinois, le kabyle, etc.). Le dialecte algérien est un langage assez particulier, un langage fondé sur un mélange de plusieurs langues dont la plupart des termes ont été repris à l'origine de la langue arabe littéraire et le français. Pour mieux illustrer certaines de ses caractéristiques dans le langage SMS, prenons cet exemple : « nakol fel restaurant » (comprendre : je mange au restaurant). L'exemple montre qu'un message écrit en dialecte algérien peut contenir un terme issu d'une langue étrangère, ici du français (restaurant). De plus, étant donné qu'il n'existe aucune règle d'écriture, un mot peut éventuellement être écrit de plusieurs manières. Par exemple, le mot «restaurant» (resto, mat3am, …) ou encore le mot «nakol» (nacol, nakoul, …) (comprendre : manger).

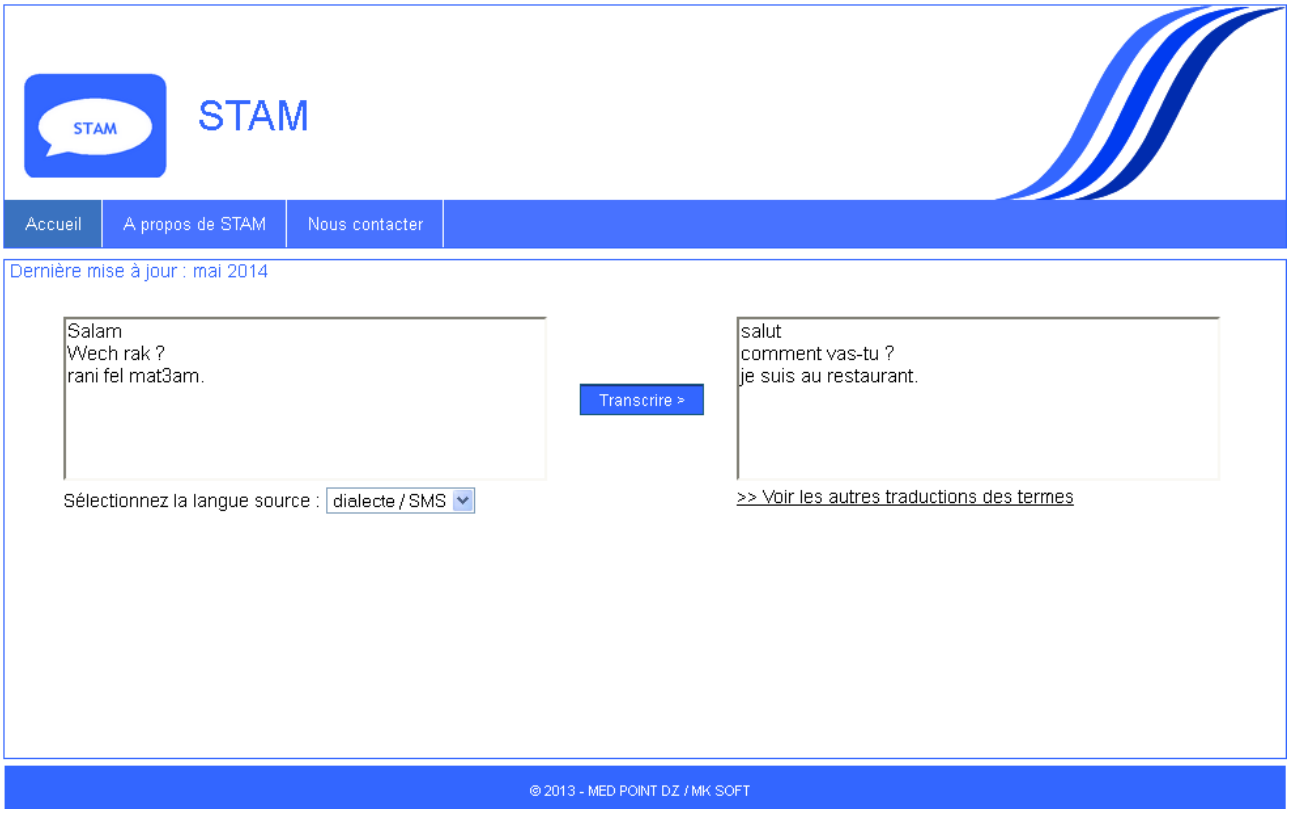
[1] Short Message Service

Une autre caractéristique du dialecte algérien (valable également pour la langue arabe) concerne la présence de chiffres dans certains mots (mat3am) pour exprimer principalement une certaine prononciation (problème phonétique) spécifique aux mots arabes.

Enfin, STAM est un système souple et facilement paramétrable permettant d'intégrer d'autres dialectes parlés dans d'autres pays. On parle ici d'une solution multi-dialectale. En effet, le but du projet STAM est de ne pas se limiter uniquement au dialecte local (algérien) mais aussi de prendre en compte, par la suite et dans la mesure du possible, les autres dialectes disponibles en commençant par les pays du Maghreb (Tunisie, Maroc, Libye, Mauritanie), puis les pays du Moyen Orient (Egypte, Syrie, Liban, Arabie Saoudite, etc.). La réalisation d'un tel outil passe par la constitution d'une importante base terminologique.

Actuellement, le système STAM s'appuie sur une terminologie multi-dialectale évolutive. Cette terminologie regroupe les termes et expressions employés dans les pays suivants : Algérie, Maroc et Tunisie. Pour les transcriptions, STAM repose également sur un ensemble de règles d'écriture et d'algorithmes. On peut citer l'algorithme « STAM_Align » qui permet d'effectuer des alignements entre la requête utilisateur et le contenu de la terminologie.

Ci-dessous une figure représentant un exemple de transcription dans le système STAM :




## Références

GUELLA N. (2011). Emprunts lexicaux dans des dialectes arabes algériens. *Synergies Monde arabe 8*, 81-88.

VANHOVE M. (1999). Les dialectes arabes des régions sud, centre et est du Yémen : perspectives et recherche. *Chroniques Yéménite*, 95-100.