JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

GENEVIÈVE LALLICH-BOIDIN

Données linguistiques et traitement des questions ouvertes

Journal de la société française de statistique, tome 142, n° 4 (2001), p. 29-36

<http://www.numdam.org/item?id=JSFS_2001__142_4_29_0>

© Société française de statistique, 2001, tous droits réservés.

L'accès aux archives de la revue « Journal de la société française de statistique » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

DONNÉES LINGUISTIQUES ET TRAITEMENT DES QUESTIONS OUVERTES

Geneviève LALLICH-BOIDIN *

RÉSUMÉ

Les réponses aux questions ouvertes dans une enquête sont des textes que l'on souhaiterait coder *a posteriori* pour pouvoir les exploiter au sein des autres réponses qui sont elles codées *a priori*. Au travers des résultats d'une enquête sur le centre-ville de Grenoble, outre les traditionnelles difficultés du traitement de la langue, se pose le problème des traces écrites résultant de la transcription des réponses. Après un examen des différents problèmes, cet article tente de proposer des solutions simples, pour résoudre les problèmes soit en amont soit en aval.

Mots clés : questions ouvertes, transcription oral-écrit, polysémie, synonymie, traitement automatique de la langue.

ABSTRACT

Answers to open-ended questions in a survey are considered as texts. One would like to be able to code them *a posteriori*, in order to process all answers at once since other answers are coded *a priori*. Through a specific inquiry about Grenoble town center, we noticed that besides the traditional difficulties of natural language processing, another problem arises due to the transcription of messages, from spoken to written form. After examination of the situation, this paper proposes simple solutions acting either before or after processing.

Keywords : open-ended questions, synonymy, polysemy, spoken to written language transcription, natural language processing.

1. Introduction

Les réflexions qui suivent sont issues de l'étude du corpus des réponses aux questions ouvertes obtenu suite à l'enquête téléphonique effectuée auprès d'un échantillon aléatoire et représentatif des habitants des communes de l'agglomération grenobloise, dans le cadre du DESS « PROGIS Etudes d'opinion et de marché ». Elles ont pour objectif de proposer des solutions aux différents problèmes qui surgissent dès lors que l'on souhaite traiter automatiquement des séquences langagières. Nous tenons ici à remercier les organisateurs de cette enquête. En effet, il n'est pas si courant que l'on puisse à

* Laboratoire RECODOC - Université Claude Bernard, 43 bd du 11 novembre 1918, 69622 Villeurbanne cedex. E-mail : genevieve.lallich-boidin@univ-lyon1.fr

la fois connaître les conditions dans lesquelles l'enquête s'est déroulée et disposer des données recueillies. Nous examinerons les conditions de recueil des données et le matériau récolté. Puis, nous nous placerons du point de vue de l'analyse automatique de la langue sur les questions ouvertes et examinerons les problèmes soulevés afin de tenter d'y apporter des solutions.

2. Les conditions du recueil des données

Des conditions dans lesquelles s'est déroulée l'enquête, nous retiendrons les faits suivants :

Tout d'abord, les enquêtes ont été menées par téléphone. L'enquêteur avait donc une double tâche à mener : poser les questions oralement, retranscrire la réponse orale de l'enquêté. La plupart des questions étaient des questions fermées, limitant ainsi le problème de la retranscription. Certaines questions, semi-ouvertes, donnaient lieu à un post-codage par l'enquêteur (profession, diplôme...). Parmi les questions réellement ouvertes nous avons traité :

Q47 : si je vous dis centre ville de Grenoble, à quoi pensez-vous spontanément ?

Cette question est posée en début d'enquête après avoir demandé où habitait l'enquêté et s'il avait le sentiment d'habiter le centre ville de Grenoble.

Q71 : si vous deviez imaginer les deux choses les plus importantes qui devraient être améliorées d'ici 20 ans dans le centre ville, quelles seraient-elles ? (consigne : bien relancer si une seule raison)

Cette question arrive après quelques questions fermées énumérant les activités propres au centre ville, ses « qualités »...

Ce sont donc les réponses (au nombre de 404) à chacune de ces questions qui seront l'objet de notre propos. Les réponses ont été retranscrites par les enquêteurs à la volée. Ont-ils tout retranscrit, ou seulement ce qu'ils jugeaient pertinent ou croyaient entendre ? Nous ne le saurons pas. Certaines réponses sont très lapidaires, d'autres beaucoup plus nuancées : on est en droit de penser qu'il y a un effet « enquêteur », effet qui est étudié par P. Caillot et M. Moine (2001) dans ce même volume.

Exemples extrêmes de réponses à la question 47 :

- *place Grenette*
- *Il a bougé. quand j'étais jeune, j'allais sur la place Grenette, maintenant, on se demande où il est. Grenoble s'est étendue comme une pieuvre il y a plusieurs centres à mon sens. Par exemple, la maison de l'architecture au niveau de la place st andré, un autre centre à victor hugo, des marchés de noel, le troisième, la villeneuve, mais je ne connais pas bien. il a bcp perdu de sa psychologie et de ce qu'il avait avant. il n'y a plus de lieu de vie, il n'y a que des grands magasins.*

Par ailleurs, la syntaxe est celle de l'oral et la graphie est dictée par l'efficacité.

Les transports ; le tram c bien, ms il va trop loin. il faut remonter sur l'gglo. aller vers montobbonot, st egrêve. Allez vers l'agglo.

Ces remarques, bien que suggérées par cette enquête, sont généralisables à toute enquête menée dans des conditions similaires. Elles nous amènent donc à nous poser la question : quels traitements automatisables peut-on effectuer sur ces données, dans quel but ? Quels obstacles faudra-t-il surmonter ?

3. Les objectifs visés

La première des questions ouvertes a pour fonction d'appréhender les différentes connotations attachées spontanément à « centre-ville de Grenoble ». L'objectif est donc de détecter les différentes dimensions associées à centre ville (lieux, activités, émotions, qualités) ainsi que les différentes valeurs. Ce que l'on vise c'est un codage *a posteriori.*

La deuxième question ouverte (les deux choses les plus importantes qui devraient être améliorées d'ici 20 ans) arrive après de nombreuses questions à choix multiples entre différents items. On s'attendrait à trouver une majorité de réponses au futur ou au conditionnel (seulement 2 réponses au futur et 18 au conditionnel). En fait la plus grande partie des réponses consiste en la reprise d'items proposés dans les réponses précédentes :

les places de parkings / la circulation en générale

augmentée parfois par des adverbes de comparaison (*plus, moins, mieux*) :

plus d'espaces verts et moins de voitures

accompagnée parfois de verbes d'évolution (*améliorer, développer, réduire, augmenter, supprimer, diminuer...*) ou de leur nominalisation :

réduction des voitures / augmenter les transports en commun

la circulation, c'est à dire la supprimer.

Ces verbes et adverbes permettent d'introduire la notion de changement par rapport à la situation actuelle.

Un traitement de ces réponses doit donc permettre de faire émerger les deux « choses » à améliorer ainsi que le sens de l'amélioration (plus, moins) pour chacune d'elles, afin de pouvoir lier ces réponses au reste du questionnaire.

4. La correction des données

4.1. Des données bruitées

Les chaînes de caractères correspondant aux réponses sont, pour de nombreuses raisons, dont celles tenant aux conditions de recueil, le résultat de variations autour de la langue. L'idée première est de corriger ces textes pour les amener vers une graphie plus standard et donc réduire l'amplitude des variations. Cette correction a été faite manuellement pour cette enquête par un des responsables de l'enquête. D'une manière générale, le correcteur est amené à se poser la question de savoir quelle est la forme à atteindre, car corriger c'est faire un pas vers le codage. Comment écrire *centre ville* (*centre-ville, Centre-ville...*) Est-ce un mot unique ou deux ? Dès lors que l'on est amené à

corriger un texte en vue de traitements automatiques, on est amené à se poser la question : quelle forme faut-il atteindre ? Car corriger, c'est interpréter.

4.2. Corrections

Il est donc nécessaire avant de corriger de se fixer le but à atteindre et donc la norme : non pas la norme absolue mais une norme liée au corpus et aux traitements visés. Et se fixer une norme, c'est se donner un lexique et une syntaxe.

Dans le cadre de cette enquête, il apparaît évident que le lexique doit contenir des formes complexes comme : centre-ville, transport(s)-en-commun, grands-boulevards... et des noms propres comme Grenoble, Trois-Tours, Ile-Verte, Place-Grenette, Place-du-Tribunal. Il s'agit de ramener toutes les variations autour de ces formes à celle choisie.

Quant à la syntaxe, elle nous vient de l'oral : il n'y a pas lieu de la transformer.

La correction automatique ne semble pas envisageable car chaque enquêteur dispose de son propre encodage. Mais une correction humaine, la seule possible, est difficilement systématique.

5. Les hypothèses sous-jacentes au traitement des réponses

Les textes réponses recueillis sont soumis à des logiciels de traitement automatique de questions ouvertes dans les enquêtes [SPHINX – M.-L. Gavard-Perret & J. Moscarola (1998), ALCESTE – M. Reinert (1996), SPAD-T – L. Lebart & A. Salem (1994)]. Ces traitements mécanisés s'effectuent sur la forme du matériau recueilli et reposent donc sur les deux hypothèses implicites suivantes :

1. l'oral et l'écrit sont deux versions d'une même langue : l'écrit est donc le reflet fidèle du message oral.

2. une réponse à une question ouverte est une production langagière très contrainte. On doit de ce fait circonscrire assez aisément les problèmes inhérents à la langue (polysémie, synonymie).

Nous examinerons tour à tour la validité de chacune de ces hypothèses.

5.1. L'écrit est-il l'image fidèle de l'oral ?

Sans entrer dans un débat fort intéressant qui anime la communauté linguistique, à savoir si l'écrit et l'oral sont deux faces d'un même système ou bien si ce sont deux systèmes distincts, nous sommes amenés à nous poser dans le cas qui nous préoccupe la question de savoir si les transcriptions écrites sont fidèles aux messages émis par les enquêtés. Car exploiter une enquête c'est analyser les réponses des enquêtés au travers de la transcription obtenue par l'enquêteur.

Le constat, que bien des linguistes ont fait avant moi, est malheureusement négatif (Anis, 1989), ceci pour deux raisons :

L'oral neutralise des oppositions que l'écrit devrait distinguer. En effet, l'écrit est plus précis que l'oral. C'est ainsi que l'oral ne marque pas de nombreuses oppositions (singulier / pluriel, marque du sujet des verbes, majuscule / minuscule) que l'écrit distingue. De ce fait, si le contexte ne permet pas d'opter pour une forme ou l'autre, l'enquêteur fait nécessairement un choix lorsqu'il écrit, et lève ainsi une indétermination. Est-ce justifié ?

Exemples : *plus de parking (ou parkings) pour se garer*

Ne sait (ou sais) pas difficile de répondre

Est-il dans ce cas bien nécessaire de distinguer, dans les comptages de formes, les singuliers des pluriels, les flexions verbales...

De façon duale, l'écrit neutralise des distinctions de l'oral. En effet, l'écrit ne rend pas compte des phénomènes de prosodie, d'accent tonique, de voyelles muettes... Le cas le plus frappant rencontré dans l'enquête est le suivant :

y'a plus de place

Le singulier indique-t-il que l'oral était de la forme ([iaplydplas] ≈ *il n'y a plus de places*) plutôt que de la forme ([iaplysdeplas] ≈ *il y a plus de places*). Seul l'enquêteur savait lequel des deux énoncés avait été émis. La trace écrite est ambiguë sauf à considérer que l'enquêteur a choisi le singulier de *place* pour rendre compte de l'énoncé [iaplydplas]. Ceci contredit notre propos du paragraphe suivant, à savoir que les marques de nombre étaient peu fiables. Sans recours à l'enregistrement, nous ne pouvons qu'émettre des hypothèses, sans pouvoir les confirmer ou les infirmer. À titre d'illustration, une intervenante dans une émission télévisée a dit [õnaplysdavãtaj... õnãnamwẽ] « *on n'a plus d'avantages... on en a moins* », prenant conscience de l'interprétation multiple de son premier énoncé.

Au travers de ces exemples fournis par l'enquête, l'on ne peut soutenir l'hypothèse qu'il y a « bijection » entre l'oral et l'écrit. Trois possibilités s'offrent pour réduire les écarts dus à la transcription de l'oral. La première consisterait à traiter la trace orale des réponses, en utilisant des logiciels de reconnaissance de la parole. On échapperait ainsi à la transcription. Il faut espérer que les variations interlocuteurs et que les conditions de recueil des enregistrements ne produisent pas plus de bruit qu'une transcription spontanée. Les performances actuelles des logiciels de dictée vocale nécessitent à la fois une période d'apprentissage pour chaque locuteur, et requièrent que celui-ci élimine expressivité, effets prosodiques et hésitations de ses productions.

Une deuxième solution consisterait à enregistrer les échanges en cours d'enquête. Cela permettrait aux enquêteurs de différer la transcription, qui n'aurait plus lieu en temps réel. Une telle solution est coûteuse car le temps de saisie d'une enquête en serait au moins doublé. Mais le principal obstacle est déontologique.

La troisième solution consisterait à demander aux enquêteurs de transcrire les réponses dans un écrit correct quitte à ajouter des flexions, des formes non formulées lorsqu'elles sont justifiées par les marques orales non transcriptibles

mais permettent de respecter le sens de l'énoncé émis. Cela demanderait une petite formation des enquêteurs pour leur indiquer les pièges principaux de la transcription et leur donner des moyens d'y remédier. Le temps de saisie en serait sûrement augmenté, mais la qualité des résultats améliorée.

Exemple :

L'énoncé [iaplydplas] se réécrira *il n'y a plus de places*, et non pas *y'a plus de places.*

Alors, *il y a plus de places*, se comprendra comme la transcription de [iaplys-deplas].

Le problème de la transcription de l'oral vers l'écrit en français a fait l'objet de nombreux travaux en linguistique, car elle est à la base de l'écriture. À ma connaissance, la transcription dans une optique de restitution de l'oral a été abordée pour le français par le Groupe Aixois de Recherche en Syntaxe (Claire Blanche-Benveniste, 1986, 1987), dont le but est de constituer des corpus écrits de langue parlée pour étudier les variations syntaxiques de l'oral en français.

5.2 Les problèmes inhérents à la langue

Rappelons la deuxième hypothèse que nous avons formulée. Une réponse à une question ouverte est produite dans un contexte très contraignant. Les enquêtes ont été réalisées dans un intervalle d'au plus une semaine. Par ailleurs, les questions ouvertes arrivent après une série de questions fermées, induisant ainsi une thématique, un vocabulaire... Seules varient les caractéristiques des personnes enquêtées. Étant donnée l'importance de ces contraintes, on serait en droit d'attendre que les phénomènes linguistiques qui perturbent la bijection entre signifiant et signifié (polysémie et synonymie) soient réduits. Or, il n'en est rien. Pour cela, nous avons étudié l'usage des formes nominales les plus fréquentes, en neutralisant les oppositions singulier / pluriel pour les raisons évoquées précédemment.

En ce qui concerne les synonymies, nous avons relevé trois exemples de poids. Le tableau ci-dessous donne le nombre d'occurrences de chacune des formes rencontrées :

automobile(s)	31	voiture(s)	86		
commerce(s)	42	boutique(s)	11	magasin(s)	67
bouchon(s)	5	embouteillages(s)	9	encombrement	1

On serait en droit d'arguer que *automobile* et *voiture* ne sont pas parfaitement synonymes. En effet, il s'avère que *automobile* est soit nom, soit adjectif dans *circulation automobile.* Mais on trouve aussi bien *moins de voitures* que *moins d'automobiles.* Une analyse morphologique automatique doit permettre de distinguer ces deux emplois. On réduira la polysémie mais on ne la radiera pas. On notera aussi que l'évocation des encombrements fait appel le plus souvent à l'un des deux termes issus d'une métaphore avec la *bouteille.*

Il en est de même pour *boutique* et *magasin* car l'usage veut que l'on préfère *petite boutique* et *grand magasin* à *grande boutique* et *petit magasin.* Les fondre en une même catégorie, relève du choix de l'enquêteur.

Il est assez simple de remédier à la synonymie dans un corpus fermé. Il suffit de choisir un représentant de chacune des classes et d'y ramener les synoymes.

Enfin, la polysémie est elle-aussi présente. L'élément lexical le plus fréquent du corpus est *place*, 300 occurrences, sous l'une des deux formes graphiques *place* et *places.* Or, coïncidence, il est lui aussi le plus polysémique. Ceci s'explique en partie par le fait qu'il a dans la langue de multiples emplois (place de parking, place de théâtre, et qu'en plus dans le contexte du centre-ville, il sert à désigner des lieux : place Victor-Hugo, place de Verdun, place Grenette). Voici quelques illustrations :

places de stationnement

plus de places pour se garer / les places de parking / pour qu'on puisse stationner quel que soit le type de places(?)

place – espace vital : *qu'il y ait moins de circulation, plus de place aux piétons qu'aux voitures*

place – espace urbain : *réaménagement place Grenette / il faudrait aménager les autres places, les espaces verts*

mettre en place (locution) : *il faudrait mettre en place davantage de manifestations sur la place Victor Hugo*

Il en est de même pour le mot *sécurité* - 42 occurrences

sécurité - dualité piéton – automobile : *la propreté et la circulation ou la sécurité...ça se rejoint / il y a tellement de voitures que tout le monde traverse au feu rouge, y'a plus de sécurité...*

sécurité - phénomène de société : *la sécurité, on peut pas sortir avec un sac à main*

Comment interpréter ? : *améliorer la sécurité*

Pourtant chacun de ces exemples relève d'un traitement différent. Quand le mot *place* détermine un espace minimum, il joue plus le rôle de déterminant comme *tas* dans *un tas de.* Il est donc le plus généralement accompagné par un spécifieur nominal *place de parking.* Lorsqu'il est suivi d'un nom propre (Grenette, Victor-Hugo, de Verdun), il dénomme un lieu, le codage peut permettre de les isoler. Il paraît donc possible en utilisant des schémas de type place de Nom, place Nom-propre, de désambiguïser un grand nombre d'occurrences du mot *place.* En revanche, il semble beaucoup plus difficile de cerner les diverses significations du mot *sécurité.* Une raison à cela : le terme *la sécurité en ville* est un des items d'une question précédente. Même si les enquêtés l'ont perçu différemment, ils se réfèrent à leur interprétation quand ils citent *la sécurité* sans préciser davantage (c'est le cas le plus fréquent). Seuls quelques enquêtés conscients de la polysémie de ce mot ont ressenti le besoin de préciser leur interprétation. Quelle est donc la valeur de ce mot quand elle n'est pas précisée ? Celle sous-entendue par l'enquête ?

Le problème de la polysémie est beaucoup plus redoutable car il est le propre de la langue. Et l'on atteint les limites du codage. Mais les questions ouvertes n'ont-elles pas aussi pour raison de permettre à l'enquêté de s'exprimer ? C'est par elles que l'on peut percevoir les multiples interprétations d'un terme que l'on pensait univoque.

6. Conclusion

Les énoncés recueillis au travers des questions ouvertes sont très riches eu égard aux contraintes qui pèsent sur eux. Cette enquête sur le centre-ville de Grenoble doit être considérée comme une illustration des problèmes généraux que l'on rencontre dès lors que l'on souhaite traiter des questions ouvertes, et non pas comme un cas particulier.

On notera, cependant, qu'un nouveau mode d'enquête émerge via le WEB : il permet aux enquêtés de s'exprimer par écrit. On élimine ainsi les distorsions dues à la transcription.

Quelle que soit la méthode d'exploitation de l'enquête employée, il restera toujours vrai que ce n'est pas le traitement de la langue qui crée des ambiguïtés, c'est la langue qui est « ambiguë » et son traitement ne peut que révéler cette caractéristique fondamentale.

Bibliographie

ANIS J. (1989), De certains marqueurs graphiques dans un modèle linguistique de l'écrit. *DRLAV-Revue de Linguistique*, 41, pp. 33-52.

BLANCHE-BENVENISTE C. et JEANJEAN C. (1986, 1e éd., 1987, 2e éd.), *Le français parlé, Transcription et édition*, Didier-Erudition/ InaLF, Paris.

CAILLOT P. et MOINE M. (2001), Mais quelle est la réponse ? Quelques problèmes posés par l'exploitation d'une question ouverte administrée par téléphone. *Journal de la Société Française de Statistique*, 142, 4, pp.

GAVARD-PERRET M. L. et MOSCAROLA J. (1998), De l'énoncé à l'énonciation : pour une relecture de l'analyse lexicale en marketing. *Recherches et applications en marketing*, vol. 13, n° 2.

LEBART L. et SALEM A. (1994), *Statistique textuelle*. Dunod, Paris.

REINERT M. (1996), Un logiciel d'analyse lexicale : ALCESTE. *Les cahiers de l'Analyse des Données*, 4, pp. 471-484.