Introducing Instruments in the Generative Lexicon

Asanee Kawtrakul, Mukda Suktarachan
asanee_naist@yahoo.com
Kasetsart univ. Bangkok, Thailand,
Sudeshna Sarkar
sudeshna@iitkgp.ac.in
IIT Kharagpur, India,
Patrick Saint-Dizier, Elixabete Murguia
stdizier@irit.fr, elimurguia@yahoo.com
Irit-Cnrs, Toulouse, France

Abstract

We show how the different conceptual facets of the instrumental role of a concrete object can be characterized and introduced in the telic role of the Qualia structure. We outline different difficulties, among which prototypicality and context dependence. Using various languages with an explicit instrumental grammatical case or with specific type restrictions does help to identify instruments among other proposition adjuncts.

1 Aims and Motivations

Instrumentality has a quite wide conceptual scope, almost anything can potentially be used as an instrument for a number of tasks, and instrumentality largely overlaps with other complex notions such as causes, paths and manners.

It is difficult to give a comprehensive definition of what instrumentality is. In WordNet it is defined as 'an artifact, or a set of artifacts, that are instrumental (i.e. behave as instruments) in accomplishing some end', i.e. reaching a certain goal. In this definition, the triple relation agent-instrument-goal (as in: *John cuts the bread with a knife*, where *John* is the agent, *knife* is the instrument that does the cutting, and *bread cut* is the goal), is left vague in what concerns the exact involvement of the agent and the instrument in the action, and the control the agent has on the instrument and on the action (Mari and St-Dizier 01). In this definition, nothing is specified about the 'deep' nature or the prototypicality of the instrument. WordNet simply lists some quite diverse types of prototypical instruments: systems, means, implements, hardware, furnishings, equipment, devices, means of transportation, container, etc.

If almost anything can be an instrument, we can nevertheless formulate a few criteria. First, an instrument is basically non volitional. When humans are used as instruments, they are not volitional by themselves. The action is controlled by another agent who takes the initiative of the action. Next, there must be a kind of control relation between the agent and the instrument (e.g. the agent has the physical ability to use the instrument).

Instrumentality has been relatively widely studied in psychology, there is much less material in linguistics, and even less in computational circles. However, the notion of instrument is important, for example, it is a key notion in Qualia telic roles and, more practically, in question-answering systems to be able to process questions about instruments.

In this paper, we consider instrumentality as conveyed by prepositions or equivalent means (e.g. postpositions, affixes). Our aim in this study is twofold:

1. to identify the conceptual facets of instrumentality so that a conceptual semantics can be defined in the spirit of (Talmy, 01), (Wierzbicka 92, 96) and
2. to elaborate an accurate enough model so that, for a given object, its most prototypical function(s) as an instrument can be introduced in an appropriate manner in its Qualia telic role.

Instruments must be modelled as relational types, involving at least agent and object, w.r.t. an action. Then within the framework of instrumental question-answering, the response of a query may be elaborated from elements from the a Generative Lexicon, instead of going through Web pages. We can then view the generative lexicon as a repository for instruments, or, even, as a set of indexes, prepared for answering instrumental questions.

Since almost any object can serve as an instrument for a number of actions and considering the large number of metaphors, we will restrict ourselves in this paper to instruments denoted by concrete objects. Since those objects may be instruments for a large number of tasks, we then need to develop a measure of prototypicality so that basically only major uses are accounted for in the Qualia telic role of this object. Obviously, it would be of much interest to be able to derive less standard uses

from the standard ones, as a kind of coercion system, but probably this is much more pragmatic than linguistic.

In our work, instruments for actions are identified from two sources: the Web, for general actions, and technical texts for specialized procedures. While with the Web we are confronted with problems of syntax (identifying PPs which are instruments) and with prototypicality (the Web abounds in unexpected uses of instruments, largely metaphorical and creative, in GL terms), technical texts give us a quite narrow view of instrumental uses of concrete objects, restricted to a few operations. A kind a synthesis between the two views seems to be necessary.

Finally, instead of focusing on a specific language, and besides the conceptual study reported in the next section, we found it useful, in order to resolve ambiguities between what is an instrumental expression or not, to develop a multilingual approach (Kawtrakul et ali. 06), considering languages from various families, in particular those with an explicit grammatical instrumental case mark (however, we could not include the study of instrumental case in Slavic languages).

2 The conceptual parameters of instrumentality

In a first stage, let us analyze in depth the different parameters of instrumentality, which will be of an operational interest for the telic role. The work reported here is based on the observation of instrumentality over 12 languages from 5 families: French, German, Spanish, Italian, Berber, Arabic, Hindi, Kasmiri, Urdu, Bengali, Thai and Malay (Kawtrakul et alii. 2006). The characteristics of these languages allow for the definition of quite accurate conceptual distinctions, from the various marks that they require. These marks include: prepositions, postpositions, affixes, morphological marks, and derived terms like verbs, meant to introduce certain types of instruments in specific contexts.

Here are, briefly sketched, the different conceptual distinctions we have identified, that seem to be useful for our present work.

2.1 Concrete instruments

All languages studied have at least one basic instrumental mark operating over concrete objects (T: duai, M: dengan, H: se, U: se, K: sity, B: diye, -e, -te, G: mit, S: con, F: avec, A: bi, BR: èg). Several refinements are identified, for specific types of NPs, or to denote a specific intention:

- the instrument is a recipient (S: en) or, more generally, conveys an idea of container (e.g. spoon) (B: -e kare), the idea behing is that the container is used to carry the object along a certain trajectory,
- the instrument is a part of the body (e.g. hand): T: kap. In this case, the instrument is not strictly artifactual.
- the goal is difficult to reach, it requires some efforts from the agent (S: a base de),
- the focus can be emphasized by using dedicated marks (G: mit Hilfe (von), Mittels (more fomal: *Das Gericht hat mittels einstweiliger Verfügung den Drogenhandel untersagt* (the court has with provisional ordinance the drug traffic prohibited))).

The second major difficulty is prototypicality (Rosch, 78). When the instrument used is not very prototypical of the action, several languages re-inforce the instrumental prepositions to, sort of, coerce the type of the noun so that it can become an acceptable instrument. We have examples in S: `por medio de`, B: `sAhAjye, sahojoge`, I: `per mezzo di`, F: `au moyen de, par le biais de` (biais= bias which directly expresses this idea), as in:
F: *Il a ouvert la porte au moyen d'un cric* (he opened the door by means of a jack).

At a conceptual level, it is quite difficult to characterize what is a prototypical instrument for a given action (characterized by subject-verb-object: *John opens the door*). Each event has its own prototypical instrument, making corpus studies extremely large, probably unfeasable. When searching on the web, we find an incredible variety of instruments to open a door, almost impossible to classify. Next, prototypicality is not a boolean notion: instruments are more or less prototypical. Since the instrument is very much dependent on the verb and on the object, we cannot foresee any form of incorporation in the verb that would give us indications. A direction could be to assume Qualia structures (Pustejovsky 91) associated with each potential instrument that describes the function of the object in the telic role. For example, key(X) would have open(X, door), with *door* being quite generic. This approach could work via a large lexical development for concrete nouns, it is much more risky when terms are abstract.

2.2 Abstract instruments

Abstract instruments (theorems, regulations, examples, etc.) are realized identically to concrete instruments, but with some typical marks such as: T: tam,

H: dwAra, K: zariyi, B: dwArA, M. mengikut. At this stage, it is difficult to explain why marks are different from concrete instruments. An hypothesis could be that abstract instruments are closer to causes (see 5.5), or to more formal situations for which specific terms were developed (e.g. for G: kraft).

There are additional marks dedicated to particular fields: B: sahajoge, and A: `min khilal` when instruments are of type 'example' (explain with an example). U: `-ke zariye`, S: `por medio de` and G: `Anhand, Kraft` are more formal, stronger for `Kraft` and apply particularly to areas like juridical or psychological domains.

People and organizations can be seen as appropriate intermediaries for reaching a goal. They may be conceived as metaphorical instruments. Investigations show that people can get controlled much in the same way as concrete objects:
F: *Elle a informé Paul de son départ par Pauline* (She informed Paul of her leaving 'by' Pauline).

If we now consider: S: *Juan envío este paquete por correo* (John sent this parcel 'by' post)
Since post is the by-default medium to send packages, `por` is the only choice. Using more precise services, like FedEx, is considered to be an alternative way, in that case F: `par, avec` S: `por, con` are both acceptable.

Let us note also that, for example in French or in Spanish, *au moyen de*, *por medio de* (by means of) are used to introduce instruments which are not prototypical. This is a way to force or coerce the 'type' of an object to be an instrument, by virtue of the preposition semantics.

2.3 The overlap instrument-manner

In a number of cases, it is not very easy to make a distinction between instrument and manner. It seems there is a continuum between these two notions or even some form of overlap, where the object is both an instrument and a manner at various degrees, which may depend on context. A variety of marks contribute to characterize this overlap, manners at stake being quite diverse, but we will not go into the study of manners. Specific marks dealing with the manner/instrument ambiguity are: T: `doi`, G: `durch` (which is also used for metaphorical spatial uses), M: `dengan menggunakan`, S: `en, con, a` as in S: *escribir en/con rojo* (write in red),
T: *khian - duai - muek - daeng* (write - with - red - ink)
BR: *tète s-éfèssen* (She-eats with-hands).

2.4 Causality

It is clear that, a priori, instruments can be viewed at various degrees as causes of an event. There is a kind of overlap between these two notions. Instruments are not volitional, so they are under the partial or full control of an agent (humans playing the role of instruments are also controlled by an agent). Typically I: `a causa di`, F: `a cause de`, S: `a causa de` signal that the instrument has brought about an event:
I : *Il castello e distrutto a causa di un violento incendio.* (The castle has been destroyed 'because of' a violent fire.)

Causality (e.g. Talmy, 01) being a complex notion, it is not surprising that instruments, viewed as intermediaries at various degrees, share some features with causes. For example in *cut the bread with a knife*, the cause of the bread being cut is the action of the agent, but also the use of a prototypical property of the knife: the knife does the cutting. In (Talmy 01), the instrument is embedded into the causing event:
(caused event) RESULTS FROM (causing event)
where the causing event has the structure:
Instrument ACT ON object, where object is bound or related in some way to the object in the caused event.

As analyzed in (Mari and Saint-Dizier, 01), instrumentality is the convergence of several factors:

- the degree of involvement of the instrument in the action, therefore, the fact that the instrument causes the action or is just a means managed by the agent who is the main cause,
- the type of control the agent has on the instrument for the action at stake, from full control to lack of control,
- the control the agent has over the action as a whole.

Indic languages and Thai are particular explicit on these matters. They have specific marks for two major cases:

1. agentive instrument, action not controlled by the agent: H: `ke dwAra`, U: `ke zariye` K: `zariy, desi`, T: `doi`,
2. causal instrument that does most of the action, under the control of the agent: H `ke kAran`, U: `ki vajah se`, K. `kiny`, T: `duai`,

Berber allows `èg` only when the agent controls the instrument. The other cases are expressed by non prepositional forms.

2.5 Instruments and paths

Another productive situation is the use of spatial metaphors to express instrumentality. The use of F: `par` and other marks (e.g. in B., U.), show that there is a close link between instrumentality and path descriptions (spatial as well as temporal paths). This is a kind of metaphorical use of paths viewed as instruments (as can be seen in (Lakoff et al. 99): 'action is motion, goals are paths, actors are travellers'). Using an instrument parallels the use of paths in the domain of space.

Marks denoting paths or sources are of much interest. Some have really restricted uses, whereas others are more flexible. We observe the following main components:

- paths: T: tam, A: min khilal, S: por, a través de(por correo, by post), de, G: durch, F: à travers, note the distinctions, e.g. in M: melalui (metaphorical paths: M : *berhubung melalui telefon* (communicate by telephone)), menerusi (channel of transmission), H, U: me, se, T: thang. In B, -e and -te denote paths where the agent that does the action has no control, whereas diye and dhare involve at least a partial control from the agent. In M, metaphorical passages require `melalui`.
- sources: F: à, A: min a, T: chak (for concrete and abstract sources). Example: A: *Achroubou mina Karoura* (I am drinking with bottle), which is also a kind of manner.

The duality path/instrument is particularly visible in, e.g.:
K: *raam vot tshochi vati kiny gari* (ram reach-past short route via home = Ram reached home by the short route).

Another interesting phenomenon occurs when an argument is both an instrument and a path, as in *look at the moon in a telescope*. Telescope is indeed the instrument used and also the path through which one looks, or which the light traverses. This double facet of the argument is visible in surface realizations, where the preposition used is ambiguous between instrument (first preposition) and path (second one) readings: G: `mit, durch`, S: `con, por`, M: `dengan, melalui`, F: `avec, dans`. When one wants to strongly stress the path interpretation, then a more path-oriented preposition is used, e.g. S: `a través de`.

2.6 Means of transportation as instruments

Means of transportation (trains, spoons, boxes, envelopes, etc.), sometimes viewed as containers, and mediums of transportation (by air) receive a special treatment in a number of languages: T: doi, thang, M: menerusi, melalui (for metaphorical mediums and passages), H: me, U: ke zariye, K: manz, zariy, B: kare, -e kare, -ya kare, A: ala, BR: ge-, kh-, G: per, S: por, en, F: par. We have, for example:
U: *raam gaadi ke zariye daftar gayaa* (ram car by means of office go-past = ram went to post office by car)
T: *pai - pa ris - doi - khrueang bin* (Go to - Paris - by - plane)
B: *Nouko-ya kare phuketa jAo* or *Nouko-ya phuketa jAo* (boat-e kare phuket go or boat-e phuket go = go by boat to Phuket)

A distinction is made between the medium and the means as for: M: secara, which is used for means of communication such as email or letters. If the agent has effective control over the means, then, for example, S uses `con`.

2.7 Our target

In order to have a quite refined instrumental description of objects in the telic role, we keep the following distinctions: concrete, abstract (possibly metaphorical), manner, causal, agentive, path (with subdivisions such as source), and means of transportation. A concrete noun can obviously have several instrumental roles of various kinds.

3 The challenge of acquiring instrumental information

Let us concentrate, in this short paper, on the acquisition of instrumental roles from Web data for concrete nouns. So far, the method is essentially manual, due to the complexity of the analysis and of the decisions to be made, which still need accurate evaluations. The current study has been realized on a sample of 100 concrete nouns from various technical domains. This study is useful prior to developing acquisition tools for instrumental functions.

3.1 Data collection

To collect data from the Web, we use a very simple bootstrapping method, which consists in using a 'seed' as a query and then collecting related data. We have two converging tests: (1) we use a well-defined instrumental mark + a concrete noun (e.g. by means of a hammer) and we collect a whole set of forms: Verb-Object-Instrument or (2) we start from a verb + the object (cut glass), expecting an instrument. The verb chosen is based on our intuition of a prototypical action to be made on the object.

These tasks are realized first on one language from the above mentioned. However, whenever necessary, this task was realized in parallel on several of the above languages, in order to have a more

accurate analysis of instruments and also to eliminate a number of terms which are misinterpreted as instruments, due to the polysemy of marks. This is realized by taking into account instrumental marks that we use to filter out PPs which are not instrumental. This requires the use of multilingual ressources in order to get appropriate translations of seed terms and responses in the various languages studied. These multilingual tests were used only when data was ambiguous.

3.2 Organizing data

The next problem is to organize the data we obtain around 'prototypical kernels'. Advanced question-answering, since 2004 (Harabagiu et ali. 2004), have investigated techniques to select or produce a single answer among a set of candidates, which is not fully coherent. This is realized via fusion techniques, applied either on numerical data (Moriceau 2006) or on conceptual data. Most algorithms are based on probability distribution measures that outline semantic approximations. A set of seven fusion operators have been elaborated, which are relevant for our purpose: contradiction (with degrees), addition, refinement, agreement, generalization, tendency and irrelevance (see (Webber et ali. 2002)).

In general, we get a variety of responses to a given seed test. For a given object, we may get variations (by refinement, agreement, generalization) around a similar prototypical function, or a number of different types of functions (by addition, tendency, possibly weak contradiction) which depend on the task. As a complement, if we now consider a given task, we may similarly get closely related instruments, rather prototypical, or a set of different types of instruments, if the task is either generic or if there are several ways of doing it. For example, for 'cut glass', the instrument to be used depends on the type of glass, on the type of object to realize (e.g. an aquarium) and e.g. on the type of edge desired.

Responses found on the Web are extremely diverse (e.g. for the above example: *cutter, diamond cutter, CNC cutter, water jet cutter, diamond point, laser, tungsten carbide wheel, diamond steel small wheel, hammer*), and some of them are really unexpected (*head, hand, heart, words, songs,* etc.). Some of these instruments are closely related to each other and are different specialized variants. For example, we have different types of *cutters, namely, cutter, diamond cutter, CNC cutter, water jet cutter,* etc. These are all variants and a cutter is kept as the most prototypical of these instruments.

More generally, in order to avoid a proliferation of possible uses of a concrete object, we need to elaborate a strategy that identifies its most prototypical functions, and the objects it applies to. Unfortunately, this may be highly contextual. The statistical distributions observed for verbs describing functions allow us to quite easily identify a prototypical verb (in the above example, for the seed 'by means of a cutter', cut and related verbs have an occurence frequency of 78%, the other verbs encountered being e.g. *clean, draw, scratch*). For a given task, instrumental objects are much more diverse. We need to identify them via a graph, realized from a conceptual metrics and a general purpose ontology or equivalent system like WordNet (Fellbaum 93). The graph characterizes conceptual distances between objects in a simple way in our case, and therefore it allows for the identification of one or more prototypical instruments, as advocated above, for question-answering (which is in fact our ultimate goal in this project).

3.3 Representing instruments in Telic roles

In the telic role, the instrumental function is therefore represented not by a single predicate as generally assumed, but by a logical expression, where:

1. the action performed is expressed by a generic verb predicate including an event notation,
2. the different circumtances, if any, are specified by means dedicated predicates which are, in our case, mainly thematic role labels that we use in our applications,
3. the resulting situation or the goal is specified by a predicate goal(event, GOAL)), where the verb that appear in GOAL specifies the purpose in a quite generic way. GOAL can be a single predicate, a formula, or any other appropriate semantic form such as a Lexical Conceptual Semantics representation (LCS). The LCS is used e.g. when there is no direct predicate to encode the resulting situation.

For example, we have, for the following nouns used as instruments the following telic roles (X denotes the object and Y the actor):
cutter(X):
telic: concrete instrumental role:
cut(X,glass,e) $\wedge$ manner(e,sharp) $\wedge$
goal(e,build(Y,acquarium)).

scissors(X):
telic: concrete instrumental role:
cut(X,hair,e) $\wedge$ goal(e, shorten(Y,hair))
cut(X,paper,e) $\wedge$ goal(e,
$[_{event}\ GO_{+char,+ident}([_{thing}\ paper],$

$[_{path}\ From_{+char,+ident}([_{property}\ one\ piece]),$
$To_{+char,+ident}([_{property}\ several\ pieces])])$

example(X):
telic: abstract instrumental role - example:
explain(X,contents,e) $\wedge$ goal(e,clarify(Y,contents))

An interesting remark is that, on the web, instrumental expressions are often associated with warnings (e.g. precautions to take) and prerequisites, which are quite informative, these could be included in a response in a question answering system.

4 Results and perspectives

This preliminary work has allowed us to:

- define a quite accurate conceptual model for instrumentality in connection with its language realizations, over a quite large set of languages. This allows us to characterize several operational categories of instruments, and predict their use in concrete situations,
- propose, via the study of about 100 seed tests, a strategy for organizing groups of instruments collected from the web via fusion techniques and a conceptual metrics. The work was carried out from two perspectives: the object and the action, and from several languages, for a better accuracy and precision,
- show how instruments can be represented by means of a logical representation in the Qualia telic role for concrete objects.

The goal is now to partly automate this procedure, so that the work can be further extended. As a side effect, we will then be able to answer instrumental questions with much more accuracy. This application was, in fact, the motivation for this more generic work.

References

Fellbaum, C., (1993), *English Verbs as Semantic Net*, Journal of Lexicography, vol. 6, Oxford University Press.

Harabagiu, S., Lacatusu., F., Strategies for Advanced Question Answering. In Proceedings of the Workshop on Pragmatics of Question Answering at HLT-NAACL, Boston, USA, 2004.

Kawtrakul, A., et alii. (12 authors): A multilingual Analysis of the notion of Instrumentality, 2006. *Proc. EACL workshop on prepositions*. Trento, Italy.

Mari, A. (2000), *Polysémie et Décidabilité. Le cas de avec ou l'association par les canaux*, Thèse de Doctorat, EHESS, Paris, collection Langue et Parole, L'Harmattan, 2003.

Moriceau, V., Generating Intelligent Numerical Answers in a Question-Answering System. 4th International Natural Language Generation Conference (INLG), p103-110, Sydney (Australia), 15-16 juillet 2006.

Saint-Dizier, P., (2005), *PrepNet: a Framework for Describing Prepositions: Preliminary Investigation Results*, IWCS05, Tilburg.

Spark-Jones, K., Boguraev, B., A note on a study of cases, research note, Journal of the ACL, dec. 85 available electronically at www.acl.org.

Talmy, L. (1976), *Semantic Causative Types*, In M. Shibatani (ed.), Syntax and Semantics 6: The Grammar of Causative Constructions. New York: Academic Press, pp. 43-116.

Talmy, L., (1985), *Lexicalization Patterns: Semantic Structure in Lexical Forms*, in *Language Typology and Syntactic Description 3: Grammatical Categories and the Lexicon*, T. Shopen (ed.), 57-149, Cambridge University Press.

Talmy L., 2001. *Towards a Cognitive Semantics*, vol. 1 and 2. MIT Press.

Webber, B., Gardent, C., Bos, J., Position statement: Inference in Question Answering. In Proceedings of LREC, 2002.

Wierzbicka, A. (1992), *Semantic Primitives and Semantic Fields*, in A. Lehrer and E.F. Kittay (eds.), Frames, Fields and Contrasts. Hillsdale: Lawrence Erlbaum Associates, pp. 208-227.