Jennifer Pearson, SALIS, Dublin City University

The Expression of Definitions in Specialised Texts: A Corpus-based Analysis

Abstract

This paper explores the notion that terms are often, either explicitly or implicitly, defined or explained in certain types of texts. I argue that defining can be described as a performative action when, to paraphrase, Austin, the circumstances are felicitous. I describe the felicity conditions which must be met in order for statements to be considered as partial definitions or explanations of terms. The work outlined in this paper forms part of a broader investigation into the possibility of producing specifications for the (semi-) automatic identification and retrieval of terminological definitions from corpora.

1. Introduction

This paper sets out to demonstrate that certain specialised text types contain a combination of language and metalanguage statements and that many metalanguage statements are in fact complete or partial definitions of terms used in texts. Our investigation is based on three corpora: the ITU corpus (4.7 million words) available on CD-ROM from the University of Edinburgh, the GCSE corpus (1 million words) held at the Cobuild Unit at the University of Birmingham and a collection of texts (229,676 words) from the Nature Journal, kindly made available to me by Tim Johns at the University of Birmingham.

2. Definitions as performatives

Austin (1962) uses the term 'performative' to describe verbs which, when used, invoke some form of conventional procedure and which in themselves constitute some form of action. These procedures have what Austin calls illocutionary force which means that the speaker, in *saying* something, is actually *doing* something. Austin (1962:65) includes the act of defining in his set of performatives: "When we use the formula, 'I define x as y' we have a transition to a performance utterance." Defining can therefore be construed as a performative action. It is suggested here that, like other categories of performatives, definitions do not have to be

817

explicitly signalled in order to be valid. They need not be preceded by expressions such as 'I hereby define..'

Authors will define the terms they use for a number of different reasons. If they have created and named a new concept, they are likely to define the term when they first introduce it. If a concept and a term already exist within a particular subject domain, an author may wish to expand or redefine the concept underlying the term, thereby altering the definition. These are what we term original defining performatives because terms are being defined or redefined; these performatives are likely to be found in academic papers written by subject specialists for their peers. When a concept and term already exist, an author may choose to reiterate the definition for the purposes of clarification. This typically occurs in texts such as those which we have in the ITU and GCSE corpora. We term these reported definitions because they involve the clarification of meaning and are therefore explanations rather than original definitions. I think it is possible to argue that both original and reported definitions are performatives in the sense defined by Austin, in that they both have illocutionary force. In this paper, we focus on reported definitions and, in particular, on formal reported definitions.

A number of conditions must be met for a statement to qualify as a reported definition. These conditions are analogous to those specified by Austin and relate to authorship, readership, text function and the syntax of the statements. With regard to authorship, readership and text function, we suggest that texts about a special subject domain which have a didactic or informative function and are written by authors with a certain level of expertise for a readership with an educational or professional need to read these texts are likely to contain reported definitions. Authors writing within this particular communicative context are likely to explain at least some of the terms which they use.

3. Retrieval of formal definitions from the corpora

We have opted to focus on the retrieval of reported definitions which match one of the following patterns: 1) $X = Y +$ distinguishing characteristic, or 2) $Y +$ distinguishing characteristic $= X$, whereby X is subordinate to Y. The formal definition will provide three types of information (adopted from Trimble, 1985:80): the name (i.e. X) of the *term* being defined, the *class* (i.e. Y) to which the term belongs and the *difference(s)* between the term and all other members of the class. Each of these types of information must be present for the statement to be considered. Contrary to what one might normally expect of formal de-

finitions, the term which is being defined can appear either before or after the defining statement. Flowerdew (1992:168) suggests that the location of x in the sentence depends on whether or not it has been introduced previously in the discourse. It will appear before the defining statement if already known and after the defining statement if it is being introduced for the first time. The evidence of our corpora would certainly support his first point but we are unable to agree with his second point because there are instances in our corpora where x appears after the defining statement, even when it has already been introduced in the discourse.

As there are many statements in the corpora which match the above patterns for formal definitions but which are not themselves formal definitions, we need to specify a number of conditions in order to retrieve only those which are actual definitions. We have specified two sets of conditions. The first relates to specifications for the slot fillers for X, Y and '='. The second set applies to the statement as a whole and consists of the conditions which a statement must meet in order to qualify as a formal definition. Each of these will be discussed separately and exemplified. We start by specifying the slot fillers for X, Y and '=' and then specify the conditions for the statement as a whole.

3.1 Specifying slot fillers for X, Y and '='

X must be a term and must match one of a specified set of term formation patterns. If X is not a term, the statement is immediately eliminated from consideration. We have also specified that if X appears before the defining statement, it must be preceded by a) the indefinite article b) no article at all. In other words, it may not occur with any modifiers other than the indefinite article. This condition allows us to ensure that the terms retrieved have generic rather than specific reference. While there are instances of formal definitions where X appears before the defining statement and is preceded by the definite article or by the demonstrative adjective, these belong to what Trimble (1985) calls complex formal definitions, i.e. statements where the definition extends beyond the sentence delimiters. These do not concern us in this paper as we are concerned only with the retrieval of simple formal definitions. If X appears after the defining statement, it may not occur with any modifiers other than the definite or the indefinite article.

Y must be either a term which matches one of a specified set of term formation patterns or one of a set of specified class words. The class words are generic terms such as *process, method, function, property*. Regardless of whether it appears before or after X, Y may not occur with

any modifiers other than the definite or the indefinite article and it may occur without any modifier at all.

A number of words may fill the '=' slot and we have chosen to call these *hinges*. A hinge is a verb or verb phrase which links X with Y. Hinges are what Sager (1980) terms *connective verbs*. In English Special Languages, Sager suggests that descriptive statements tend to be very simple, "consisting merely of two nominals linked by verbs such as *to be, to have* or *to give*." (1980:186.) The connective verbs which he lists include: *effect, assure, perform, obtain, provide, involve, entail, imply, decrease, increase, restrict, limit, result in, lead to, be ascribed to, be attributed to, result from, amount to, measure*. We took the list of verbs provided by Sager and examined our corpora to establish 1) whether they occurred in the corpora and 2) whether they were used in the manner described by him. We found that many of these verbs hardly ever occurred in our corpora; for example, *entail* occurs only once in the GCSE corpus, *imply* occurs only four times in the same corpus and *is ascribed to* never occurs in the GCSE corpus. When these verbs do occur, they do not appear to function in the way that Sager has suggested. We decided therefore to devise our own set of hinges or connective verbs. Our list is based on textual evidence. We distinguish between two classes of hinge, class 1 and class 2 hinges. With class 1 hinges, X appears on the lhs and Y + the distinguishing characteristic appear on the rhs; with class 2 hinges the order is reversed. Class 1 hinges include *comprise(s), consist(s) of, define(s), denote(s), describe(s), designate(s), is/are.* Class 2 hinges include *is/are, is/are called, is/are known as.*

3.2 Specifications relating to the statement as a whole

In addition to specifying the slot fillers for X, Y and '=', we have specified a further set of conditions relating to the statement as a whole. The statement 'x = y' must constitute the main clause of the sentence and may not be prefaced by clauses or phrases other than those which fill the *x, y* and '=' slots. This allows us to exclude sentence-initial clauses which might affect the general applicability of a statement.

The hinge must be in the present tense, indicative mood. It may be in active or passive voice. The use of the present tense gives the statement general applicability; this would not be possible were the hinge in a past or future tense. Hinges which occur with modals are excluded from consideration. Modals restrict the scope of a statement and introduce an element of doubt. If X *can be defined as* Y + distinguishing characteristic, we are unable to infer that it is always defined as such. The use of the

modal suggests that other definitions are possible which means that the statement does not have universal applicability. Hinges which are qualified by negating words such as *not, never* do not qualify for consideration. We are interested in discovering what terms are rather than what they are not.

Focusing adverbs (e.g. *generally, usually, commonly*) are used quite frequently in the GCSE, ITU and Nature corpora. The function of these adverbs is to restrict the scope of the reference. What is interesting is that, in normal circumstances, one would consider that the use of some of these adverbs with any of the hinges would suggest that the statement was generally applicable and therefore had some form of definitional status. In fact, the reverse is true. Focusing adverbs prevent statements from having generic reference. If a statement which meets all of the other essential criteria is modified by one of these adverbs, it does not qualify as a potential definitional statement. When a term is described as *usually* having a particular characteristic, it is not possible to conclude that it *always* has this characteristic.

Y must be followed by one of the following: a preposition, a relative pronoun, a past participle. It may not be followed by the co-ordinating conjunctions *and, or, but*. If *Y* is followed immediately by *and* this may signal the introduction of another noun phrase or the beginning of another clause rather than the introduction of the distinguishing characteristic. If it is followed immediately by *or*, this is likely to signal the introduction of another noun phrase. The use of *but* tends to signal an exception rather than the rule.

4. Analysis of statements

All of the examples provided here meet the criteria specified for formal definitions. To facilitate analysis, we classified the examples according to the type of hinge used and, for each hinge, according to the class of word used to introduce the distinguishing characteristic. As previously indicated, two broad classes of hinge were used for the analysis of formal definition statements. In the first class, *X* appears on the lhs of the hinge and the defining statement on the rhs while, in the second class, the order is reversed. Section 4.1 contains examples of class 1 hinges (*is/are, consist(s) of, is/are defined as*) where *Y* is followed either by a past participle or a relative pronoun; examples of each of these are provided. Section 4.2 contains examples of the class 2 hinge, *is/are called*, where *Y* is followed either by a past participle, a relative pronoun or a preposition. Examples from the ITU, GCSE and Nature corpora are prefaced by A, B

and C respectively. One example of each pattern is provided. When no example is provided, this indicates that the pattern does not occur.

4.1 Class 1 hinges

4.1.1 *X is/are Y* + distinguishing characteristic

A. A misdelivered frame is a frame transferred from a source user to a destination user other than the intended destination user.

B. Bile is a green liquid made in your liver and it is stored in your gall bladder.

C. INTERLEUKIN-1 (IL-1) is a cytokine1 produced by mononuclear phagocytes.

A. An expedited-data-unit is a service-data-unit which is transferred and/or processed with priority over normal service-data-units.

B. Vorticella is a unicellular animal which lives in ponds, puddles and sewage filters.

C. Kinesin is a motor protein that uses energy derived from ATP hydrolysis to move organelles along microtubules.

4.1.2 *X consist(s) of Y* + distinguishing characteristic

This particular hinge is interesting because one would normally expect it to introduce a list of component parts rather than a superordinate term. While it is indeed used to introduce lists in our corpora, it is also used in the same way as the *is/are* hinge. It is easy to distinguish between the two uses because when it is used to introduce a list, *Y* is either modified by a number or followed by the conjunction *and*. When *Y* is not modified by a number and/or not followed by the conjunction *and*, we know that *consist(s) of* is being used to introduce a superordinate. In the Nature corpus, this particular hinge is not used to introduce formal definitions of the type described here and is used only to introduce part-whole relations which explains why no examples are provided.

A. Direct parameter input consists of an optional parameter block entry sequence preceded by the separator : (colon).

B. Skeletal muscle consists of bundles of muscle fibres held together by connective tissue.

A. Plain language consists of words that present an intelligible mean-ing in one or more of the languages admitted for international tele-

grams, which include at least French, English and Spanish.

B.   A nephron consists of a cup-shaped, hollow Bowman's capsule (Fig. 22.5) which leads into a long, narrow tubule.

4.1.3 *X is/are defined as Y* + distinguishing characteristic

A.   A confirmed service is defined as a service which results in an explicit confirmation from the service-provider.

C.   A unit of inhibitory activity is defined as the amount that produces the same level of inhibition of the binding of 1ng 1251-labelled L-1 alpha protein to El4-6.1 cells as 1ng recombinant IL-1 alpha protein.

4.2 Class 2 hinges

4.2.1 *Y* + distinguishing characteristic *is/are called X*

A.   A message described by a probe is called a described message.

B.   The black substance made from the remains is called humus.

A.   A functional object that provides one link in the MTS store-and-forward chain is called a message transfer agent (MTA).

B.   A material that is made up of only one type of atom is called an element.

A.   A connection between an incoming and an outgoing circuit at interfaces to other exchanges/networks is called a transit connection.

B.   The rush of water up the beach following each wave-break is called the swash.

5. Conclusion

While all of the corpora contain simple formal definition statements, these statements are much more common in the ITU and GCSE corpora than in the Nature corpus. Given that the texts in the Nature corpus are written by experts for their peers, this is what we had expected. The simple formal definitions provided in the Nature corpus are of highly technical terms and the language used in the definitions is not readily accessible to the non-expert. The language used in the simple formal definitions provided in the ITU corpus is again fairly technical and there is an assumption that readers will already be familiar with the field.

However, simple formal definitions are much more common than in the Nature corpus, and this is again what we had expected. The ITU corpus was written for people who already have some knowledge of the field of telecommunications but need to learn more about it. Simple formal definitions are very common in the GCSE corpus and it appears that very many terms are defined. The language used in the simple formal definitions in the GCSE corpus is accessible to the ordinary reader and requires no prior knowledge of the subject.

The restrictions which we have specified allow us to retrieve all simple formal definitions. The exclusion of modals and focusing adverbs in particular ensure that only those statements which have general validity are retrieved. While the use of some modals, e.g. *can* and *may*, may not always affect the general validity of a simple formal definition, we have concluded that they introduce an element of doubt which precludes us from considering statements which contain them. Our investigation has shown that it is possible to use corpora to retrieve at least some of the definitional information which would normally be collected through consultation with subject experts.

References

Austin, J.L. 1962. *How to do things with words*. Oxford: Oxford University Press.

Daille, B. 1994. *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. PhD Thesis. Université Paris VII.

Flowerdew, J. 1992. "Salience in the Performance of One Speech Act: The Case of Definitions" in *Discourse Processes* 15. 165–181.

Nkwenti-Azeh, B. 1992. *Positional and Combinational Characteristics of Satellite Communications Terms*. Final Report, Eurotra Project. UK-CCL-UMIST.

Sager, J., Dungworth, D., McDonald, P.F. 1980. *English Special Languages: Principles and Practice in Science and Technology*. Wiesbaden: Brandstetter Verlag.

Sinclair, J.M. 1987. *Looking Up*. London and Glasgow: Collins ELT.

Trimble, L. 1985. *English for Science and Technology: A Discourse Approach*. Cambridge: Cambridge University Press.

Yang, H.Z. 1986. "A new technique for identifying scientific and technical terms and describing science texts" in *Literary and Linguistic Computing*, Vol. 1, No.14. Oxford: Oxford University Press. 93–103.