CATAPHORIC LEXICALISATION

Robert Michael Foster
Research Institute in Information and Language Processing, Wolverhampton University
Wulfruna Street, Wolverhampton, WV1 1LY, U.K.

Keywords: Controlled Language, Lexicalization.

Abstract: A traditional lexicalization analysis of a word looks backwards in time, describing each change in the word's form and usage patterns from actuation (creation) to the present day. I suggest that this traditional view of lexicalization can be labelled Anaphoric Lexicalization to reflect its focus upon what has passed. A corresponding forward-looking process can then be envisaged called Cataphoric Lexicalization. Applying Cataphoric Lexicalization to an existing phrase from a sub-language generates a series of possible lexemes that might represent the target phrase in the future. The method is illustrated using a domain specific example. The conclusion suggests that rigorous application of Cataphoric Lexicalization by a community would in time result in a naturally evolved controlled lexicon.

1 INTRODUCTION

Usually when a new concept or method is introduced into a company or community, a phrase incorporating words that are already familiar, is used to describe the new concept or method. The literature (Brinton and Traugott, 2005) gives an extensive survey of theories about Lexicalization and Language change, which suggest that identifiable processes are at work as language has changed in past centuries. I propose that a company might derive significant benefits by deliberately applying the identified processes of language change as their corporate sub language (Grishman and Kitteridge 1986) accommodates new concepts and techniques.

This paper presents a new methodology, called Cataphoric Lexicalisation, for applying this idea. A company who apply this method as an integral part of their processes for writing rules and policies, controls the actuation of new words (Weinreich 1968) as their corporate sub-language evolves (Kirby 2007). Such control will eventually result in the natural evolution of a controlled (Goyvaerts 1996) version of their corporate sub-language.

The UK company featured in this study applies a system of Authorisation codes which when printed upon a signed certificate, signify achievement of a specified level of responsibility. The creation of new Authorisation codes is a significant process in the overall development of the community's sub-language. This paper illustrates the application of the new methodology by describing the creation of a new Authorisation code.

Section 2 describes the context of the study. Section 3 describes how Cataphoric Lexicalization is different from traditional (Anaphoric) Lexicalisation. Section 4 describes application of the method and section 5 lists the generated lexemes selected by domain experts. Section 6 has conclusions and descriptions of future work, including plans for a software implementation.

2 CONTEXT

2.1 Authorisation Codes

The UK Company featured in this study operates a system of authorisation certificates issued to staff following successful completion of training and assessment courses. The system of certificates and associated reminder reports is managed using an in-house computer system. Several new authorisation codes are added to the system each year and some of these have become assimilated as new words into the lexicon of the company's sub-language (Grishman and Kitteridge 1986).

Tight formatting restrictions are applied by the computer system to the textual composition of the

Michael Foster R. (2010).
CATAPHORIC LEXICALISATION.
In *Proceedings of the 12th International Conference on Enterprise Information Systems - Information Systems Analysis and Specification*, pages 526-529
DOI: 10.5220/0002980005260529

entry into the on-screen ‘authorisation code’ field in the computer system. Each code is restricted to lexemes consisting of a maximum of 7 Upper case or 9 mixed case characters. These restrictions guard against clipping of the codes when they are printed. If these restrictions were not applied then there is a risk of ambiguous specifications of permitted duties appearing on authorisation certificates and/or lists that remind managers of the need for refresher training or re-authorisation interviews.

2.2 Mechanical Tree Harvesting

Power lines and other vulnerable components of a national electricity distribution network can be damaged during unusual weather conditions, when trees grow towards the lines and breach specified clearance distances. The featured company has a progressive policy for removing such vegetation, called Mechanical Tree Harvesting, which specifies the safe technique for removal and subsequent disposal of the vegetation as fuel for a carbon neutral biomass plant.

Staff who carry out this activity have received training in accordance with guidelines from the Forestry commission to ensure the work is carried out correctly. The company intends to issue these staff with a signed certificate so they can prove that they have been authorised to carry out this work. This paper describes how Cataphoric Lexicalization was used to create the Authorisation code.

3 CATAPHORIC LEXICALIZATION

3.1 Theory

Many Lexicalization Theories appear in the literature seeking to analyse and explain language change. A thorough review, which also lists the processes applied within the method, is provided in the publication ‘Lexicalization and Language Change’ (Brinton and Traugott, 2005).

Subsequent studies (including Okazaki et al. 2006, Gries, 2006, Baker and Brew 2008, and Cook and Stevenson 2009) have applied these language change theories in the analysis of historical documents and managed corpora. In this paper these applications of Lexicalisation theory are labelled ‘Anaphoric’ to reflect their backwards references to Language change events that occurred before the application of Lexicalisation theory.

In this first ever application of a formal method of language change to the featured domain, the aim is to accelerate naturally occurring language change processes. The language change event will happen after the application of Lexicalisation theory. Since the process involves looking forwards at possible lexemes that might reasonably represent the Target Noun Phrase (TNP), this application of Lexicalisation theories is given the label Cataphoric Lexicalisation.

3.2 Process

The steps of the Cataphoric |Lexicalisation process are as follows:

Step 1: Define the Target Noun Phrase (TNP). Make a clear statement of the Noun Phrase for which a new lexeme is required and identify other Noun Phrases previously used to refer to the concept.

Step 2: Define Corpus Boundary Criteria. The corpus that defines the sub language within which the new lexeme will be used must be clearly defined.

Step 3: Rank Language Change Processes. Apply Anaphoric (traditional) Lexicalisation processes to the defined corpus, including previous versions of the included documents, in order to identify previous language change events within the corpus. Then rank the language change events in order of frequency of occurrence. The selection of available processes quoted here is taken from the book ‘Lexicalization and Language change’ (Brinton and Traugott, 2005), however Cataphoric Lexicalization is not restricted to using only these processes.

(a) Compounding
(b) Derivation
(c) Conversion
(d) Clipping / Elipsis
(e) Blending
(f) Back Formation
(g) Initialism (acronym)
(h) Coinage (root creation)

One ranking method might be to count every occurrence of every lexeme that has resulted from each language change process and use this simple frequency count as the ranking criterion. An alternative approach might be to count just once each different lexeme formed from the application of each language change process. Cataphoric Lexicalization makes no restriction upon ranking of the language change processes. The possibility exists

for the method to be refined by introducing more responsive feature calculations (Cook and Stevenson 2009). See 'future work'.

Step 4: Apply the Identified Language Change Processes to the TNP. Perhaps the language change processes are applied to the TNP one at a time, but there will be occasions when a combination of processes is more appropriate. This is where the ranking of observed processes within the corpus can save time.

Step 5: Apply Domain Specific Evaluation Processes and Choose the most Appropriate Lexeme. Such evaluation processes will depend upon the domain but they are likely to involve domain experts.

4 EXPERIMENT

An illustrative, manually implemented experiment has been included here to demonstrate Cataphoric Lexicalization in action. The aim is to generate a range of possible alternative signifiers for a new authorisation code that represents the trained skills for the activity of Mechanical Tree Harvesting. Domain specialists will then make a selection from the available alternatives.

Step 1: Define the TNP. The definitions identified during initial discussions with domain experts are listed below:

Mechanical Tree Harvesting (p):
Present participle for Process
Mechanical Tree Harvester (n):
Sense1: Operator authorised to apply the process
Sense2: Machine used to carry out the process.

The phrase 'Tree Trimming' was identified as referring to a similar concept to the TNP.

Step 2: Define Corpus Boundary Criteria. The target community is the group of people who will interface with the UK company in relation to protection of assets through tree trimming. Therefore the corpus is compiled from all documents in the company policy library that contain the signifier 'Tree Trimming'.

Step 3: Rank Language Change Processes. No Anaphoric Lexicalization analysis results exist for the defined corpus and resource allocations do not permit a full lexicalization survey, therefore a manual ranking process is used in this case. Analysis of the existing bank of Authorisation codes revealed a clear preference for Initialism word creation in this domain.

Step 4: Apply the Identified Language Change Processes to the TNP. Some examples that were generated using the listed language change processes are provided below:

(a) **Compounding:** eg TreeHarvesting
(b) **Derivation:** eg Trestology …. Harvestism
(c) **Conversion:** eg To Trevest …. To Trest….
(d) **Clipping / Elipsis:**eg MechVesting; MTHarv
(e) **Blending:** eg. MchTHrv… MTreHrv….
(f) **Back Formation:** eg Mecher…Trester
(g) **Initialism (acronym):** eg MTHV; MTH
(h) **Coinage (creation):** eg. TreVest; Tresting

Step 5: Apply Domain Specific Evaluation Processes and choose the most Appropriate Lexeme. This step involves domain experts assigning a usability score to the candidate lexemes. The results are given in the next section.

5 RESULTS

The candidate Authorisation codes that were given a usability score of five or greater are quoted below along with their score:

Candidate Code	Usability Score
MTHarv	9
MTHV	8
RMTrestO	5

The above results highlight an observation identified during step 3. It is clear that members of this community have a preference for the initialism (creation of acronyms) language change process. However it must also be noted that an even higher score than for the acronym MTHV was awarded to a combination of initialism with head noun clipping. In accordance with the recursive possibilities offered by Cataphoric Lexicalisation, this observation will be incorporated into future applications of the process to this domain.

6 CONCLUSIONS AND FUTURE WORK

In this paper the Cataphoric Lexicalization approach to language change has been described and then applied manually. The result is a list of candidate lexeme signifiers for the Target Noun Phrase (TNP).

A pragmatic, domain specific evaluation of the output from the system was then applied, providing a more refined language change process ranking for the featured domain in addition to a lexeme that yielded the highest usability score.

A survey of available software that might be adapted to implement Cataphoric Lexicalisation had to include the default development platform for the featured UK company which is Microsoft Access (MSAccess). The Python based Natural Language ToolKit (Loper and Bird 2002) was also examined along with various commercially available software packages, including Random Word Generator (Wittmeyer 2009) from Gammadyne software.

Following this review of the available software options, the decision has been taken to establish the specified TNP without using a special tool, and then to approach the corpus analysis and language change process identification (Steps 2 and 3) using the Natural Language ToolKit (Loper and Bird 2002). Users will then be able to view outcomes from various string concatenation functions of step 4 and subsequent evaluation step 5 using an interface that will be produced using MSAccess. The hope is that the development effort will be less than might be imagined thanks to the clearly bounded policy document library, protected by a well organized change management system.

In line with observations included earlier, the rankings of language change processes in implementations of step 3 will be refined after each application of the method. Consideration will also be given to including selected Grammaticalisation theories into step 4 in order to further widen the range of generated candidate lexemes. Perhaps future versions of the system will also generate suggested ranking of generated signifiers in Step 5 by making comparisons with signifiers of the same category within the corpus or with previous selections made by domain experts.

REFERENCES

Baker, K. and Brew, C. 2008.*"Statistical identification of English loanwords in Korean using automatically generated training data."* In Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08), pages 1159–1163, Marrakech.

Brinton, L. L., Traugott, C.G, 2005 *"Lexicalisation and Language Change"* Cambridge University Press

Cook P. , and Stevenson S. 2009 *"Automatically Identifying the Source Words of Lexical Blends in English"* Computational Linguistics Volume 36, Number 1

Goyvaerts, P. 1996 *"Controlled English, curse or blessing? A user perspective'* Proceedings of the 1st International workshop on Controlled Language Applications (CLAW '96) (Leuven). 137-42

Gries, S. Th. 2006. *"Cognitive determinants of subtractive word-formation processes: A corpus-based perspective."* Cognitive Linguistics, 17(4):535–558.

Grishman, R. and Kitteridge R. (eds) 1986 *"Analys-ing Language in restricted domains: Sublanguage description and processing"* Hillsdale NJ: Lawrence Erlbaum Associates

Kirby, S., 2007. *"The Evolution of Meaning-space Structure through Iterated Learning"* In Lyon, C., Nehaniv, C., and Cangelosi, A., editors, Emergence of Communication and Language, pages 253–268. Springer Verlag.

Loper E.and Bird S., 2002. *"NLTK: The Natural Language Toolkit."* In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, pages 62–69. Somerset, NJ: Association for Computational Linguistics. http://arXiv.org/abs/cs/0205028.

Okazaki, N. and Ananiadou S., 2006. *"A term recognition approach to acronym recognition."* In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (Coling-ACL 2006), pages 643–650, Sydney.

Weinreich, U., Labov, W., and Herzog, M.I., 1968 *"Empirical foundations for a theory of language change."* In W.P. Lehman and Yakov Malkiel, (eds.) Directions for Historical Linguistics 97-195 Austin and London: University of texas Press.

Wittmeyer G. © 2009 Elemental Communications Ltd. All rights reserved. Company Registration Number: 04192923, VAT Number: GB 893 5571 78 http://www.gammadyne.com/rndword.htm