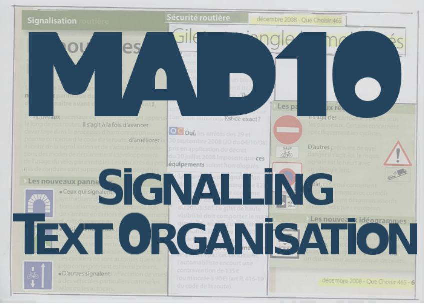
Proceedings of the 8th MAD Multidisciplinary Perspectives on Signalling Text Organisation

Lydia-Mai Ho-Dac, Julie Lemarié, Marie-Paule Péry-Woodley, Marianne Vergez-Couret

The CLLE institute (University of Toulouse) presents

Proceedings of the 8th MAD

Multidisciplinary Perspectives on Signalling Text Organisation



March 17-20, 2010, Moissac,
France

Editors:
Lydia-Mai Ho-Dac, Julie Lemarié, Marie-Paule Péry-Woodley, Marianne Vergez-Couret
CLLE, University of Toulouse-Le Mirail

Contents

1 Multidisciplinary Perspectives on Signalling Text Organisation

Multidisciplinary Approaches to Discourse 2010 is the eighth in a series of small-scale, high-quality workshops that have been organised (approx.) every second year since 1995. Its aim is to bring together researchers from different disciplines (linguistics, computational linguistics, psycholinguistics, educational and cognitive psychology, ergonomics and document design, semiotics, information and communication sciences, typography, etc.) to exchange information and learn from each other on a common topic of investigation.

1.1 Call for Papers

Signalling text organisation refers to the observation that within texts, certain features or elements seem to have a special instructional role with regard to text organisation. These text organisation signalling devices have been described under a variety of names: *signals*, *structure indicators*, *advance organisers*, *discourse markers*, *layout properties*, *surface structure features*, *organisational cues*, *stylistic writing devices* and so on. Their scope ranges from a very local level to a more global one. Their nature is also very diverse:

- *linguistic*: words (e.g. connectives), phrases (e.g. emphasis phrases), sentences (meta-discourse sentences) and beyond (overviews, summaries);
- *graphic*: typographical and spatial variation (e.g. paragraph breaks, boldface);
- *hybrid* (e.g. enumerations, headings, tables of contents, links and pop-up windows in electronic documents);
- *more elusively*: first mention, length or repetition of particular text units, structural parallelism.

Different disciplines have taken an interest in these devices, either as a core object of study or as an element to be taken into account. As a consequence, research concerned with the signalling of text organisation is far from constituting a unified field. The notion of signal itself may be associated with different key concepts according to discipline and models: *document structure*, *discourse organisation*, *layout structure*, *text architecture*, etc. As far as function is concerned, they may be seen as discourse construction devices, traces of metalinguistic segments, as reading or processing instructions, as traces of the writer's cognitive processes, or as cues revealing the author's intentions, etc.

Since the 1970's, research into the signalling of text organisation has produced considerable results. The environment for this research is at present undergoing a twofold transformation: firstly, *new methods* are appearing, linked to technological advances (corpus linguistics, natural language processing, eye movement recording techniques for the analysis of cognitive processes during reading, etc.); secondly, *new fields of application* are opening (in connection with the expanding use of digital documents in the professional and educational worlds). In this new context, novel research questions open up, requiring the integration of contributions from different disciplines or fields of study.
The aim of this workshop is to bring together researchers from different disciplines interested in the signalling of text organisation (e.g., linguists, computational linguists, psycholinguists, educational/cognitive psychologists, ergonomists, document designers, semioticians, information scientists, etc.) to allow exchanges, interactions and cross-fertilization.
We invite contributions on topics and questions such as the following (the list may be extended):

- What are text signals and what role do they play?
  - reader's viewpoint
  - writer's viewpoint
  - analyst's viewpoint
- What may be relevant theoretical models and methods of data collection and analysis to study the signalling of text organisation and its cognitive effects?
  - naturalist approaches and corpus studies
  - empirical approaches
  - micro vs. macro approaches
  - inter- and pluridisciplinary approaches
- Text signals and literacy
- Text signals in document design, natural language processing and language technologies

We welcome different types of contributions: literature reviews, theoretical and methodological considerations, reports of empirical data, corpus based-studies, etc.

1.2 Keynote speakers

John Bateman University of Bremen, Germany
Eric Hermann and Christophe Pimm CFH, Conseil en Facteurs Humains, Grenoble, France
Robert F. Lorch University of Kentucky, USA

1.3 Organising Committee

Franck Amadieu, CLLE, University of Toulouse-Le Mirail
Anne Le Draoulec, CLLE, University of Toulouse-Le Mirail

Karine Duvignau, CLLE, University of Toulouse-Le Mirail
Lydia-Mai Ho-Dac, VALIBEL, Catholic University of Louvain (Louvain-la-Neuve)
Julie Lemarié, CLLE, University of Toulouse-Le Mirail
Marie-Paule Péry-Woodley, CLLE, University of Toulouse-Le Mirail
Marianne Vergez-Couret, CLLE, University of Toulouse-Le Mirail

1.4 Scientific Committee

Franck Amadieu, CLLE, University of Toulouse-Le Mirail, France
Thierry Baccino, LUTIN, University of Nice-Sophia Antipolis, France
John Bateman, University of Bremen, Germany
Stéphane Caro, University of Bourgogne, France
Michel Charolles, LATTICE, University of Paris 3, France
Liesbeth Degand, VALIBEL, Catholic University of Louvain (Louvain-la-Neuve), Belgium
Anne Le Draoulec, CLLE, University of Toulouse-Le Mirail, France
Karine Duvignau, CLLE, University of Toulouse-Le Mirail, France
Claudine Garcia-Debanc, IUFM, University of Toulouse-Le Mirail, France
Nicolas Hernandez, LINA, University of Nantes, France
Lydia-Mai Ho-Dac, VALIBEL, Catholic University of Louvain (Louvain-la-Neuve), Belgium
Marie-Paule Jacques, LILPA, University of Strasbourg, France
Julia Lavid, Universidad Complutense de Madrid, Spain
Julie Lemarié, CLLE, University of Toulouse-Le Mirail, France
Robert F. Lorch, University of Kentucky, USA
Nadine Lucas, GREYC, University of Caen, France
Clara Mancini, Open University, UK
Fabrice Maurel, University of Caen, France
Bonnie Meyer, The Pennsylvania State University, USA
Marie-Paule Péry-Woodley, CLLE, University of Toulouse-Le Mirail, France
Josette Rebeyrolle, CLLE, University of Toulouse-Le Mirail, France
Wilbert Spooren, VU University Amsterdam, Netherlands
Patrice Terrier, CLLE, University of Toulouse-Le Mirail, France
Marianne Vergez-Couret, CLLE,, University of Toulouse-Le Mirail, France

1.5 Acknowledgements

We would like to express our thanks to our partners and sponsors for their valuable support: CLLE Institute, University of Toulouse-Le Mirail, MSHS-T (Maison des Sciences de l'Homme et de la Société de Toulouse), Département Sciences du Langage, Département d'Études du Monde Anglophone, UFR de Psychologie, Département de Psychologie Cognitive et Ergonomie, CNRS (Centre National de la Recherche Scientifique), CFH (Conseils Facteurs Humains), Région Midi-Pyrénées.

2 MAD2010 - Papers

Is the signalling of text organisation a transmodal phenomenon? — some considerations in and around language

John A. Bateman
Bremen University, Bremen, Germany
bateman@uni-bremen.de

Keywords. Discourse, Multimodality, Cues, Signals, Text Organisation

As discussed at various places in the literature, texts can be considered according to how much they rely on the visual aspects of their presentational forms in order to get their message(s) across. Bernhardt (1985), for example, describes this as a trade-off between the visual presentation modes and the verbal as suggested in Figure 1: at one end of the continuum, we have documents that come close to a traditional linguistic understanding of 'pure text', with little exploitation of the possibilities offered by visual layout and differentiation; at the other end, we find documents where more extensive use of visual possibilities is regularly made. Texts are then seen as being more or less 'visually informative'.

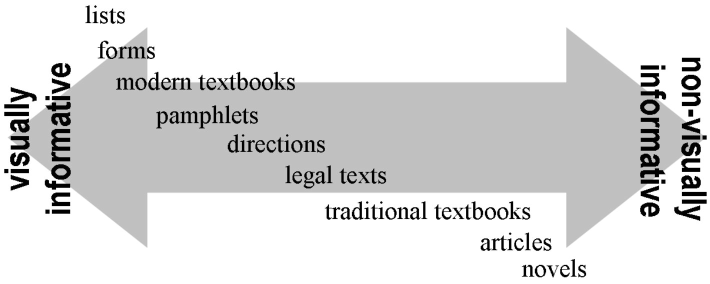
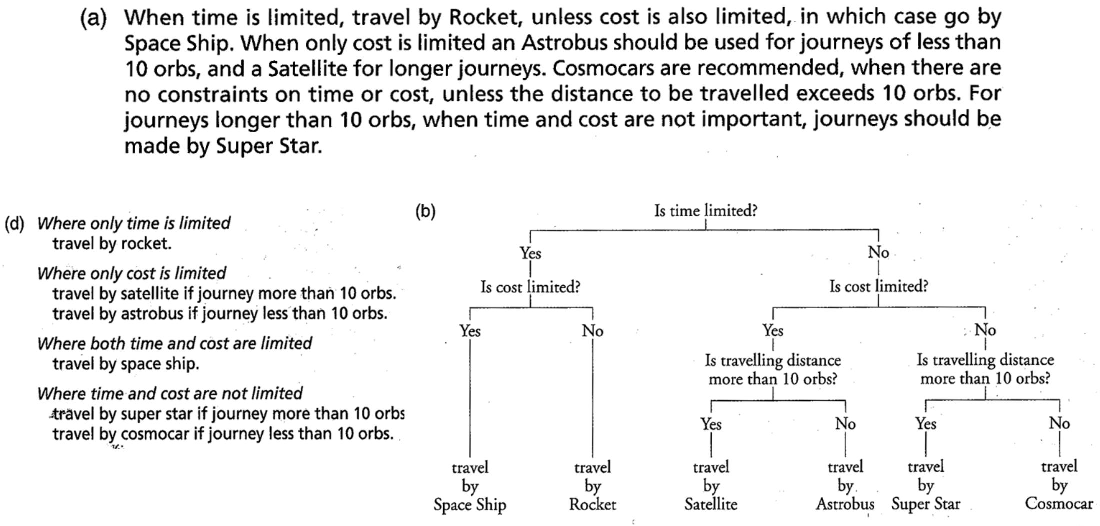


Figure 1: A continuum of visual-textual deployment adapted from Bernhardt (1985, p20)

A further set of questions that this naturally raises is just what 'information' is being presented visually, or not, as the case may be. Here, particularly in approaches starting from linguistics (cf. Lyons 1977; Crystal 1979; Matthiessen 2007), we commonly find the view that the visual representation is 'paralinguistic', adding information that might in spoken language be rendered by intonation, gesture, etc. Punctuation is here considered a borderline case: sometimes more linguistic, sometimes less (cf. Nunberg 1990).

There are, however, several other areas of 'information' that offer candidates for visual expression—and, of particular relevance here, one of these is text organisation. The visual expression of text organisation has already been approached from a variety of perspectives. In the field of document design, for example, there has been much attention paid to how arguments and other informational offerings can be presented more clearly by positively applying visual resources such as page layout (cf. Waller 1987*b*; Schriver 1997). Visual organisation, particularly segmentation realised typographically, has also been seen to provide the means for presenting what Waller has termed *access structures* into the content of documents—i.e., the ability to locate



(a) When time is limited, travel by Rocket, unless cost is also limited, in which case go by Space Ship. When only cost is limited an Astrobus should be used for journeys of less than 10 orbs, and a Satellite for longer journeys. Cosmocars are recommended, when there are no constraints on time or cost, unless the distance to be travelled exceeds 10 orbs. For journeys longer than 10 orbs, when time and cost are not important, journeys should be made by Super Star.

(d) *Where only time is limited*
travel by rocket.
Where only cost is limited
travel by satellite if journey more than 10 orbs.
travel by astrobus if journey less than 10 orbs.
Where both time and cost are limited
travel by space ship.
Where time and cost are not limited
travel by super star if journey more than 10 orbs
travel by cosmocar if journey less than 10 orbs.

Figure 2: Examples of variations across presentational modes taken from Patricia Wright

particular content quickly via visual means (Waller 1980, 1987*a*, 1990). And, particularly in computational approaches, the use of formal representations of rhetorical organisation and other semantic relations for structuring multimodal presentations has constructed an explicit link between text organisation and visual presentation (cf. André and Rist 1993; Bouayad-Agha et al. 2000; Bateman et al. 2001; Geurts et al. 2001).

In the oral presentation accompanying this brief orienting statement, I will consider several aspects of the interplay between the signalling of text organisation and the presence of visual traces of the text. These visual traces will also be taken to include the written form of language itself, i.e., the 'words on the page', not just visual or spatially presented information, so as to explore signalling options more generally. This is necessary because, here too, there appears to be a trade-off, or continuum, between the deployment of linguistic signals for text organisation, such as lexicalised conjunctions or other discourse connectives, and visual segmentation of various kinds.

One straightforward example of this, again from the area of document design, is shown in Figure 2. This shows three options for expression proposed in early work from Patricia Wright (cited as an example of the 'typographic' contribution by Crystal 1997, p193). Options (a) and (d) show variations in punctuation and macro-punctuation of the kind studied in more linguistic-based research (cf. Power et al. 2003); option (b) opens up the space of possibilities by employing more or less conventionalised diagrammatic representations of aspects of text organisation—in this case, that of flowcharts.

It is then interesting to consider the further moves taken up in the example shown in Figure 3. Here, the first is a page of instructions of how to use a video recorder discussed by Carroll and Delin (1998) and analysed multimodally further in Bateman and Delin (2003), the second is a breakfast cereal packet used as an example of typography by Crystal (1997, p197), and the third is a diagram used in a physics scientific paper in *Nature* discussed by Lemke (1998, p106). We see significant use made of the spatial layout of the text and other information in all of these examples, and so all three can be seen as 'visually informative' in the sense of Bernhardt above.

But we also see different kinds of text organisation being signalled. In the case of the video recorder instructions, the organisation is relatively clear; for the graphic from *Nature* the organ-

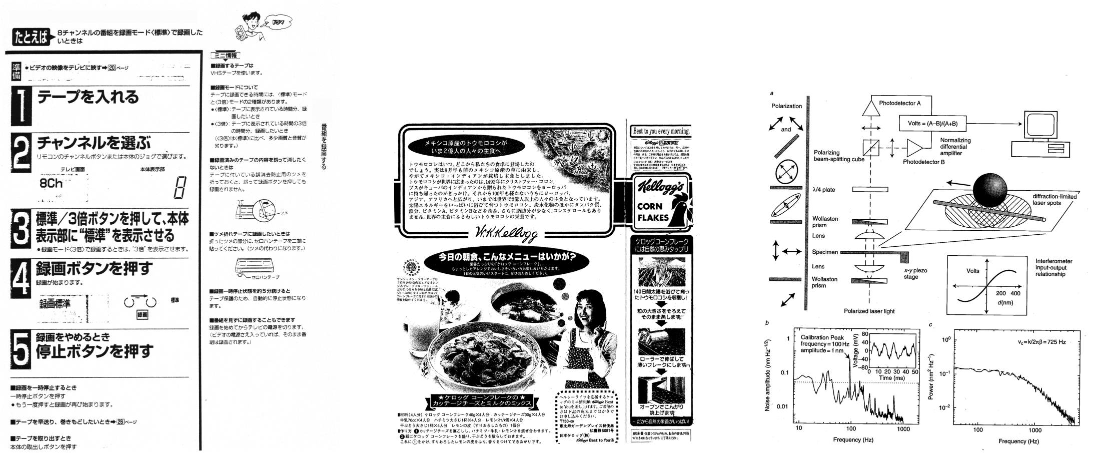

1 テープを入れる
2 チャンネルを選ぶ
3 標準／3倍ボタンを押して、本体表示部に"標準"を表示させる
4 録画ボタンを押す
5 録画をやめるとき 停止ボタンを押す
今日の朝食、こんなメニューはいかが？
Kellogg's
CORN FLAKES
Best to you every morning
Photodetector A
Polarizing beam-splitting cube
Photodetector B
Normalizing differential amplifier
diffraction-limited laser spots
Wollaston prism
Lens
Specimen
x-y piezo stage
Polarized laser light
Interferometer input-output relationship
Calibration Peak frequency = 100 Hz amplitude = 1 nm
Frequency (Hz)

Figure 3: Further use of visual/spatial presentational modes for text organisation: video player instructions, breakfast cereal and scientific diagrams

isation is considerably less clear unless one can follow the experiment being described; finally, for the cereal packet it is not at all clear without further analysis precisely what organisation is being cued, although organisation is clearly there.

In current approaches to characterising text organisation, we increasingly find accounts that abstract away from the specific modality being considered. For example, we have classifications of the kinds of text organisation that can be signalled taken from treatments of discourse connectives in verbal language being applied to static and dynamic visual representations in van Leeuwen (2005), classifications taken from rhetoric being applied across text and graphics in advertisements (McQuarrie and Mick 2003), and several more. One question to be taken up further is therefore the extent to which this is justifiable. How much do the resources available for signalling text organisation depend on the particular modalities being used? Can we take the models developed from linguistics for applications to information presentation more generally? How *is* text organisation signalled when lexicalised discourse connectives are not at hand?

Another question that it is interesting to consider is the role of the visual trace more fundamentally. To what extent are the complex text organisations that we now take for granted possible *without* written support? Clearly, complex texts can be constructed purely orally appealing to a broad range of discourse mechanisms: but to what extent do these overlap with the textual organisations constructed in written language? What models, mental and otherwise, need to be constructed by the reader/hearer/perceiver to follow the communication? And what role does the availability of the 'text-as-product' in the written mode play for our willingness to push text organisation in ever more complex directions? In this latter respect, the move to visual presentations that combine text and other modalities into complex messages, although already omnipresent, demands far closer attention in terms of its potential for expanding the possibilities available for signalling text organisation.

References

André, E. and Rist, T. (1993), The design of illustrated documents as a planning task, *in* M. T. Maybury, ed., 'Intelligent Multimedia Interfaces', AAAI Press/The MIT Press, Menlo Park (CA), Cambridge (MA), London (England), pp. 94–116.

Bateman, J. A. and Delin, J. L. (2003), Genre and multimodality: expanding the context for comparison across languages, *in* D. Willems, B. Defrancq, T. Colleman and D. Noël, eds, 'Contrastive analysis in language: identifying linguistic units of comparison', Palgrave Macmillan, Houndsmill, pp. 230–266.

Bateman, J. A., Kamps, T., Kleinz, J. and Reichenberger, K. (2001), 'Constructive text, diagram and layout generation for information presentation: the DArt$_{bio}$ system', *Computational Linguistics* **27**(3), 409–449.

Bernhardt, S. (1985), Text structure and graphic design: the visible design, *in* J. D. Benson and W. S. Greaves, eds, 'Systemic Perspectives on Discourse, Volume 1', Ablex, Norwood, New Jersey, pp. 18–38.

Bouayad-Agha, N., Scott, D. and Power, R. (2000), 'Integrating content and style in documents: a case study of patient information leaflets', *Information Design Journal* **9**(2), 161–176.
URL: *http://www.itri.bton.ac.uk/projects/iconoclast/Papers/ITRI.pdf*

Carroll, T. and Delin, J. L. (1998), 'Written Instructions in Japanese and English', *Pragmatics* **8**(3), 339–385.

Crystal, D. (1979), Reading, grammar and the line, *in* D. Thackray, ed., 'Growth in reading', Ward Lock, London.

Crystal, D. (1997), *Cambridge Encyclopedia of Language*, 2nd. edition edn, Cambridge University Press, Cambridge.

Geurts, J., van Ossenbruggen, J. and Hardman, L. (2001), Application-Specific Constraints for Multimedia Presentation Generation, *in* 'Proceedings of the International Conference on Multimedia Modeling 2001 (MMM01)', CWI, Amsterdam, The Netherlands, pp. 247–266.
URL: *http://www.cwi.nl/ media/publications/mmm01.pdf*

Lemke, J. L. (1998), Multiplying meaning: visual and verbal semiotics in scientific text, *in* J. Martin and R. Veel, eds, 'Reading science: critical and functional perspectives on discourses of science', Routledge, London, pp. 87–113.

Lyons, J. (1977), *Semantics*, Cambridge University Press, Cambridge.

Matthiessen, C. M. (2007), The multimodal page: a systemic functional exploration, *in* T. D. Royce and W. L. Bowcher, eds, 'New Directions in the Analysis of Multimodal Discourse', Lawrence Erlbaum Associates, pp. 1–62.

McQuarrie, E. and Mick, D. (2003), The contribution of semiotic and rhetorical perspectives to the explanation of visual persuasion in advertising, *in* 'Persuasive imagery. A consumer response perspective', Lawrence Erlbaum Associates, Mahwah, NJ, pp. 191–221.

Nunberg, G. (1990), *The Linguistics of Punctuation*, number 18 *in* 'CSLI Lecture Notes', Center for the Study of Language and Information, Stanford.

Power, R., Scott, D. and Bouayad-Agha, N. (2003), 'Document Structure', *Computational Linguistics* **29**(2), 211–260.

Schriver, K. A. (1997), *Dynamics in document design: creating texts for readers*, John Wiley and Sons, New York.

van Leeuwen, T. (2005), *Introducing social semiotics*, Routledge, London.

Waller, R. (1980), Graphic aspects of complex texts: typography as macro-punctuation, *in* P. A. Kolers, M. E. Wrolstad and H. Bouma, eds, 'Processing of Visible Language', Vol. 2, Plenum, New York and London, pp. 241–253.

Waller, R. (1987*a*), The typographical contribution to language: towards a model of typographic genres and their underlying structures, PhD thesis, Department of Typography and Graphic Communication, University of Reading, Reading, U.K.
URL: *http://www.robwaller.org/RobWaller_thesis87.pdf*

Waller, R. (1987*b*), Using typography to structure arguments: a critical analysis of some examples, *in* D. Jonassen, ed., 'The Technology of Text', Vol. 2, Educational Technology Publications, Englewood Cliffs, NJ, pp. 105–125.

Waller, R. (1990), Typography and discourse, *in* R. Barr, ed., 'Handbook of reading research', Vol. II, Longman, London, pp. 341–380. Reprinted 1996; Erlbaum (Mahwah, NJ).

MAD 2010, Moissac, 17–20 march 2010

Towards a pattern language approach to document description

Robert Waller, Judy Delin

The Simplification Centre, University of Reading, UK
r.waller@reading.ac.uk
judy.delin@roedelin.com

Abstract Pattern libraries, originating in architecture, are a common way to share design solutions in interaction design and software engineering. This paper introduces the approach, and explores its application to functional texts as a way of documenting common design problems along with their potential solutions. In particular, it seeks to place patterns in the context of genres, with each potentially belonging to a 'home genre' in which it originates and to which it makes an implicit intertextual reference intended to produce a particular reader response in the form of a reading strategy or interpretative stance.

Keywords: pattern, genre, layout, typography

1 Background and context

Information design is a relatively young discipline, which struggles with the lack of a usable descriptive framework. By usable, we mean one that can be used to teach or define effective design strategies that at present tend only to be known tacitly by experts. Examples of practical uses for a descriptive framework are when government regulators prescribe formats for consumer information, when publishers specify formats for textbooks, or when insurance companies set up standard styles for customer communications. So by a *usable* descriptive framework we mean, in effect, one that is to a degree prescriptive as well as descriptive.

Without anything analogous to 'grammaticality' to use as a yardstick, information design tends to rely instead on success measures that are harder to test, such as usability. In practice, rigorous testing with users is often impractical – and so practitioners rely more on 'knowing what works' from experience. Communicating this expertise, however, is not straightforward when no established descriptive framework is in place to distinguish between good and bad practice. Prescription, then, might allow us a means of judging – or at least some rules of thumb – through which we can be of use as trainers and designers in the practical world. And, in addition, the much newer field of corpus-based research on multimodal documents also lacks metrics for choosing what we should include, and what exclude, from corpora.

A number of frameworks have been proposed from within the study of typographic or information design (eg, Twyman 1979) but most aim only to be descriptive, classifying objects of analysis according to theoretical schemata. This is fine as far as it goes, but while these frameworks may help us to organise phenomena that we find, and understand the influences that underlie them, they are not intended to be the basis of the kind of practical guidance that we have argued is needed.

The programme described in this paper builds on genre-based approaches (eg, Bateman 2008; Delin, Bateman and Allen, 2002; Waller 1987, 1990), using the concept of pattern languages and pattern libraries – an approach that originates in architecture, but which has been fruitful in information design's close neighbour, interaction design. However, while genre theorists have tended to focus on explaining discourse types that already have names, the compilation of a pattern library is to a large extent a naming exercise. As one commentator put it:

> '[Naming] is one of the real powers of … patterns. They not only expose a solution but they give it a name. They create a classification system. They form a vocabulary, a language. They provide a way for people to talk about the concept and a way to recognize the solution when a similar problem context arises in the future.' Scott (2006).

2 The origin of the pattern language approach

In this context, *pattern* refers not to repeating decorative effects (for example, on wallpaper), but to configurations found consistently within recurring design solutions to common problems. They may be patterns of words, visual configurations, or a combination of both.

The term *language* needs qualifying also – it is used loosely here and does not refer just to verbal language or discourse, but to any systematic relationship between elements of almost any kind. We use it by way of reference to its originator, the architect Christopher Alexander, and in practical applications the more realistic term *pattern library* has become more common.

Christopher Alexander (1977, 1979) developed his pattern language to describe consistently observed solutions to common problems that he and his team found in a wide range of human settlements – it is a way of describing forms found in vernacular architecture that have evolved naturally in response to human needs, rather than out of theoretical models (and in particular modernist approaches).

The idea of patterns is fundamental to human thought, and is not, of course, original to Alexander. In communication theory, the definition of rhetorical patterns goes back to classical Greece, and the term is frequently used by linguists working at various different levels of analysis – particularly at the discourse level (eg, Hoey 1983, Hunston & Francis 2000). For information designers, Alexander's pattern language approach is attractive because it lends itself to a prototypical rather than taxonomic approach, which corresponds closely to how design is traditionally taught and practised (but not necessarily articulated). Experienced practitioners of any art, trade or craft are often able to recognise problems they have met before, and to call on a repertoire of possible solutions. Pattern libraries are an attempt to make explicit these traditionally tacit repertoires, and require the involvement of 'reflective practitioners' (Schön 1983) as well as descriptive analysts and this is reflected in our project team.

A typical example of an Alexander pattern is COURTYARDS WHICH LIVE (pattern 115). A courtyard allows us to resolve our desire to be outdoors and our need for protection – what Alexander calls a 'living courtyard' includes paths that cross, an opening to a wider space and a sheltered porch. Without these things, the courtyard becomes claustrophobic, rarely visited, and neglected – a 'dead courtyard'. Good spaces, created in this way, aim to achieve a quality which, having rejected as inadequate such terms as 'alive', 'whole', 'comfortable', 'free', 'exact', 'egoless', and 'eternal', Alexander calls 'the quality which has no name'. Linguists might similarly reject terms such as 'grammatical' as only partially adequate to capture the qualities of a discourse segment that it is correctly formed, relevant, cohesive and so on – and which might therefore count as a 'good' discourse contribution.

In order to help us to build in this way, Alexander captures the characteristics of what he observes to be successful environments through a series of 253 patterns (Alexander 1977). The patterns are presented systematically, and it is this approach that has been taken up in fields outside architecture – in particular, by software engineers (Gamma et al 1994). In fact, while the idea of pattern language is little more than a footnote in its original context of architecture, it is now a mainstream approach in software engineering. Software engineers were attracted to the approach because they needed a way to organise a range, or library, of configurations for software objects, to make them accessible for engineers in need of a solution to a problem that another engineer might have previously encountered.

Interaction designers (eg Tidwell 1997, 2005) have also adopted this approach. In contrast to paper document users who are expected to spend long enough with each document to become used to its unique conventions, web users move quickly between different information environments and need them to behave consistently. So interaction design as a field has had to quickly evolve a consistent set of rules that developers can use, and that users can intuitively grasp, to ensure that user effort is focused on accessing content rather than figuring out functionality. Pattern libraries have proved to be a useful way for interaction designers to share best practice.

So for Alexander, and for followers in other disciplines, a pattern is a format for capturing insight into common problems and their solutions, and for understanding the relationships between higher and lower order patterns (from a city to a shelf). In this paper we consider whether it is also a useful format for capturing similar insight about documents.

3 How problems relate to solutions in pattern languages

In his book *A Pattern Language*, Alexander describes patterns thus:

> ‘The elements of this language are entities called patterns. Each pattern describes a problem that occurs over and over again in our environment, and then describes the core of a solution to that problem, in such a way that you can use this solution a million times over, without ever doing it the same way twice.’ (Alexander *et al*, 1977:x).

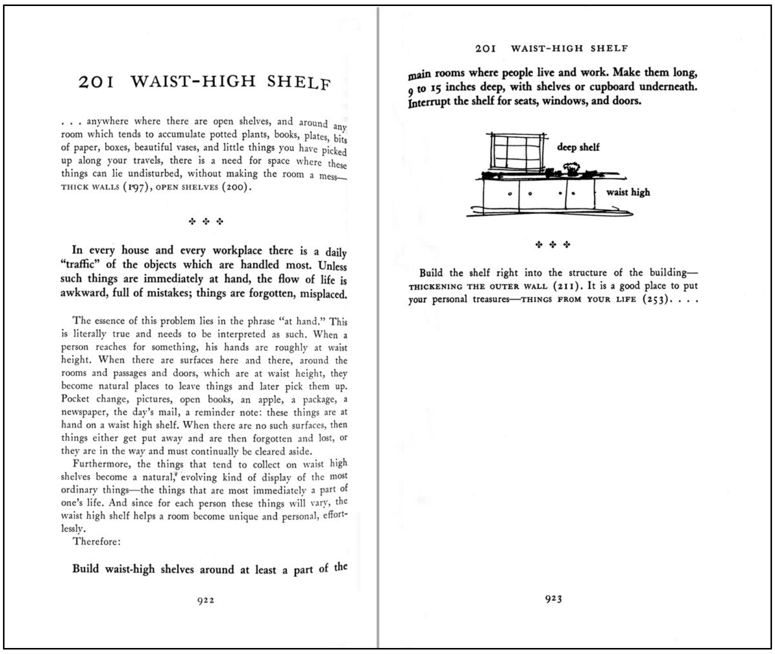
We will take as examples two of Alexander’s patterns: the ENTRANCE ROOM (Pattern 130; Alexander *et al* 1977: 622) and the WAIST-HIGH SHELF (Pattern 201; Alexander 1977:922).

The patterns take the form of the statement of a problem or need, followed by a solution. In the case of WAIST-HIGH SHELF, the problem statement is as follows:

> ‘In every house and every workplace there is a daily ‘traffic’ of objects which are handled most. Unless such things are immediately at hand, the flow of life is awkward, full of mistakes; things are forgotten, misplaced.’

There is then a discussion of how the problem might be solved, followed by a summary (in bold) of the solution:

> ‘Build waist-high shelves around at last part of the main rooms where people live and work. Make them long, 9 to 15 inches deep, with shelves or cupboard underneath. Interrupt the shelf for seats, windows, and doors.’



201 WAIST-HIGH SHELF

. . . anywhere where there are open shelves, and around any room which tends to accumulate potted plants, books, plates, bits of paper, boxes, beautiful vases, and little things you have picked up along your travels, there is a need for space where these things can lie undisturbed, without making the room a mess—THICK WALLS (197), OPEN SHELVES (200).

✣ ✣ ✣

In every house and every workplace there is a daily “traffic” of the objects which are handled most. Unless such things are immediately at hand, the flow of life is awkward, full of mistakes; things are forgotten, misplaced.

The essence of this problem lies in the phrase “at hand.” This is literally true and needs to be interpreted as such. When a person reaches for something, his hands are roughly at waist height. When there are surfaces here and there, around the rooms and passages and doors, which are at waist height, they become natural places to leave things and later pick them up. Pocket change, pictures, open books, an apple, a package, a newspaper, the day’s mail, a reminder note: these things are at hand on a waist high shelf. When there are no such surfaces, then things either get put away and are then forgotten and lost, or they are in the way and must continually be cleared aside.

Furthermore, the things that tend to collect on waist high shelves become a natural, evolving kind of display of the most ordinary things—the things that are most immediately a part of one’s life. And since for each person these things will vary, the waist high shelf helps a room become unique and personal, effortlessly.

Therefore:

Build waist-high shelves around at least a part of the

922

201 WAIST-HIGH SHELF

main rooms where people live and work. Make them long, 9 to 15 inches deep, with shelves or cupboard underneath. Interrupt the shelf for seats, windows, and doors.

✣ ✣ ✣

Build the shelf right into the structure of the building—THICKENING THE OUTER WALL (211). It is a good place to put your personal treasures—THINGS FROM YOUR LIFE (253). . . .

923

Figure 1: A typical spread from ‘A pattern language’ shows the key components: a title, an introduction that links to higher order patterns, a problem statement in bold, and an illustrated explanation.

There are two interesting things to note. One is that the pattern name in this case is the name of the solution – build a waist-high shelf. However, in the case of ENTRANCE ROOM, the solution is a particular design of entrance room, and the pattern name is more a general topic. Other pattern names are different again: SLEEPING IN PUBLIC, for example, recommends building outdoor environments that contain sheltered benches, away from traffic, where people can read the paper and doze off. In this case, the pattern named after a habit or desirable activity.

The one thing Alexander does not do is name the pattern after the problem: we don't see 'INHUMAN SCALE BUILDING', for example, or 'EVERYDAY OBJECTS OUT OF REACH'. This might be a clue to which part of the several pages that make up the pattern 'definition' is the actual 'pattern': an alternative description of pattern might be, it seems, 'loosely-specified design solution that solves a particular problem'. In the description given at the beginning, too, Alexander *et al* do suggest that patterns are both the problem *and* the solution, together.

The other interesting issue is that the pattern ENTRANCE ROOM actually contains a recommendation that there should be a waist-high shelf within the room. As point 4 of a 6-point series of recommendations, Alexander *et al* (p 624) suggest that there should be a 'shelf near the entrance' which is 'at about waist height', and provides further onward references to these and other patterns that are relevant to the satisfactory construction of the entrance room. This tells us that patterns, in his view, are recursive: it is quite normal for a pattern to contain 'calls' to several other patterns that are required to fulfil it. Whether this is full recursion or not we are not sure, but it does at least mean that patterns can be embedded within one another, in that the solution to one problem can invoke another pattern.

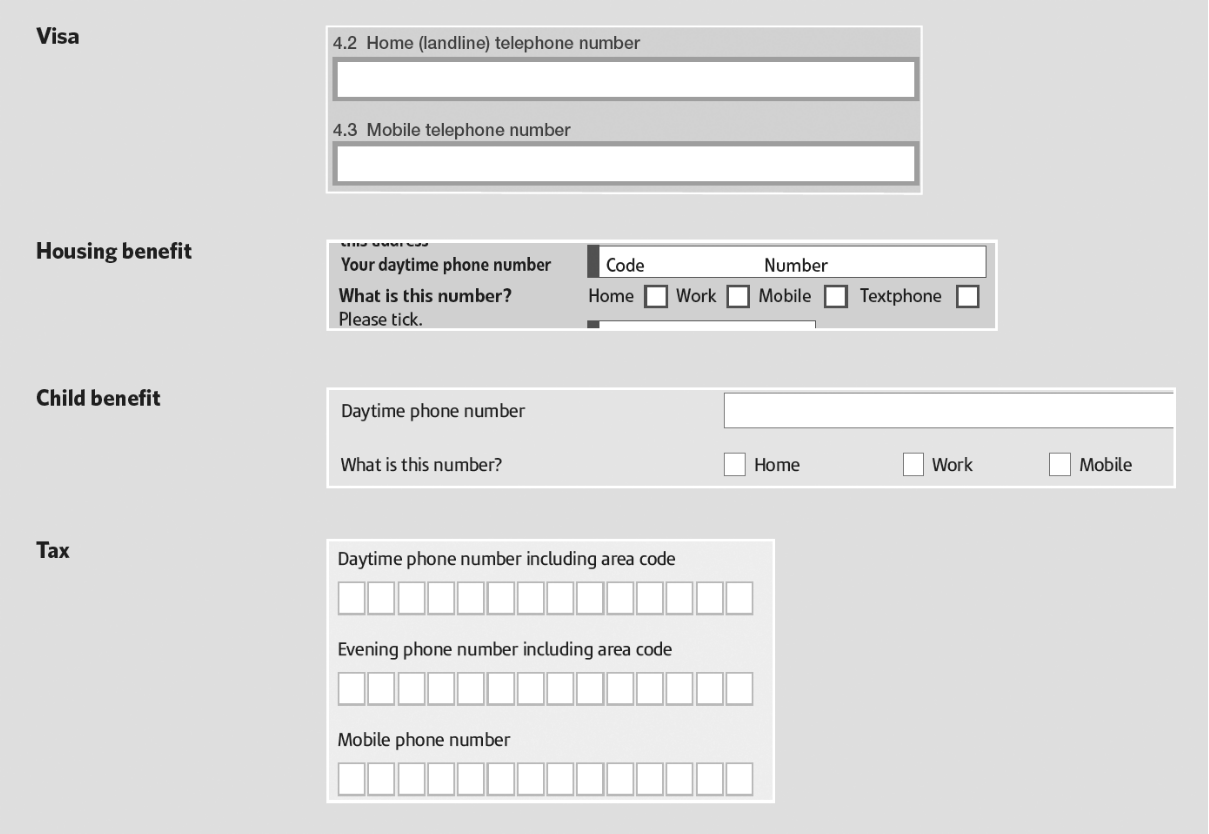
This makes sense, if we remember that problems can always be broken down into sub-problems, or goals into sub-goals, in computer planning terms. So, the problem 'make coffee' creates a sub-goal 'find coffee jar' which itself creates other sub-goals involving opening cupboards, and so on.

Going back to the summary of patterns as 'loosely-specified', it is clear from Alexander *et al*'s book that they must be so: if we 'can use this solution a million times over, without ever doing it the same way twice', there must be enough leeway in the solution to implement it in many different ways. There will be a big gap, then, between the notion of a pattern as intended by Alexander *et al*, and a notion of pattern that is implementable and computationally tractable.

4 Patterns in information design

To see whether patterns are a notion that is relevant to information design, we can look at an example of a relatively common problem in forms design: that of getting people to supply their phone number.

In Figure 2 we can see from the data collected by Crofts (2009) that there are a variety of ways of doing this even in a limited sample of four application forms. What is interesting about them is that they are more or less strongly constrained in terms of the format of the information the user can put into them.



Visa

4.2 Home (landline) telephone number

4.3 Mobile telephone number

Housing benefit

this address

Your daytime phone number Code Number

What is this number? Home □ Work □ Mobile □ Textphone □

Please tick.

Child benefit

Daytime phone number

What is this number? □ Home □ Work □ Mobile

Tax

Daytime phone number including area code

□□□□□□□□□□□□□□

Evening phone number including area code

□□□□□□□□□□□□□□

Mobile phone number

□□□□□□□□□□□□□□

Figure 2. Data from Crofts, K. (2009)

The tax form is the most constrained, in that it requires a separation of the digits into individual boxes and assumes a maximum of 14 digits. Housing benefit is the next most constrained, in that it divides the box into ‘Code’ and ‘Number’. The Visa and Child Benefit boxes are hardly constrained at all, in that they don’t suggest a format for the number or a maximum number of digits although they do employ different strategies for capturing what kind of phone number has been supplied.

We can see from this brief survey of solutions that not many people are ‘doing it the same way twice’. Some of the differences between solutions may not matter – they may be arbitrary side-effects of choices made at a different stage in the design process (for example, the choice of typeface or colour, and the thickness of lines around boxes). But some may matter in particular circumstances. For example, separate character boxes are often a sign that Optical character recognition (OCR) is being used to read the user’s data. Captioned sections (‘code’ + ‘number’) may be intended to prevent people missing out one part of the information requested. So a pattern definition needs to distinguish between its essential, or constituent features, as distinct from those that can remain accidental or contingent on other design imperatives (which might include features essential to a higher order pattern).

A question for the analyst is: looking at these samples, should we identify one loosely specified pattern, to be called PHONE NUMBER (after all, these are all reasonable ways of getting a phone number), with range of potential realisations as graphic elements, or should we identify three patterns (OPEN BOX, STRUCTURED BOX and OCR BOX), each of which has been applied to the topic of phone number, as distinct from, say, name, date or national insurance number?

5 How do patterns relate to genres?

Multi-modal studies of discourse have used genre as a key concept. Whatever else a genre may be, and however it is defined, it tends to be something that has already been given a name by its community of users: for example, leaflet, form, textbook, workshop manual, romantic novel, or crime novel. One of us has previously suggested that genre names evolve naturally, the arrival of a name signifying the achievement of communicative force by a new genre (Waller, 1987, page 285).

As we have already remarked, the identification of patterns is in one key respect the opposite of this – it is a deliberate naming exercise that recognises the existence of structures in documents that recur and are judged to be effective, but which have not acquired names naturally, except perhaps within a restricted community of practice (for example, within a particular studio, designers might refer to a layout where all items on a spread hang down from a common position, as a 'washing line', a term not shared by their readers). Pattern libraries articulate common solutions that designers use, so they can be shared and discussed.

Patterns are also distinct from genres because they are assumed to occur at various different levels of analysis, and many occur across multiple genres (that is, in documents which have very different purposes, content, format, context, etc). This was an explicit goal of Jenifer Tidwell, one of those responsible for introducing the pattern language concept to interaction design. Indeed, she saw pattern libraries as harnessing techniques not only from multiple genres but from multiple channels:

> '[A pattern language] would enable us to more methodically draw on expertise in related fields, such as book design, consumer electronics, the design of control panels (for cars, airplanes, power plants), video games, the Web and hypertext, and speech-driven interfaces.' *(*Tidwell 1999*)*

A further distinction is that while the power of genres lies mostly in their adherence to convention, patterns may work not because they represent visual conventions that readers have learned, but because they represent other sources of communicative power. For example, they may represent good 'gestalt' – layouts that communicate connections, structures and separations by harnessing the natural tendencies of our perceptual systems to seek sense in visual form. Or they may work because they represent insight into the strategies and behaviours of typical readers.

6 Prototypes and peripheries

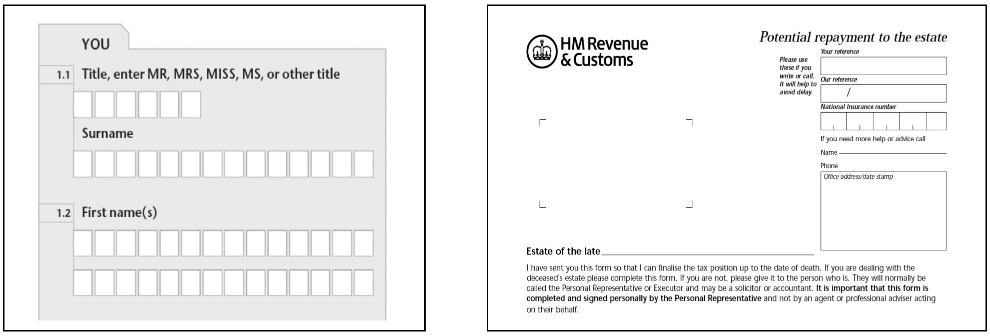
If there are some patterns that are most used, most familiar, or more constrained, or that are otherwise considered 'best' for a particular genre, we might think of those patterns as the prototypical elements of a genres. And similarly, those typographic and graphic solutions to the display of a pattern that normally work best can be thought of as prototypical solutions to a pattern.

The notion of prototype is inspired by Wittgenstein's concept of family resemblances (Wittgenstein 1953) and developed by Rosch (1973; see also Taylor 2003). It accounts for the fact that humans tend to group things into classes for the purposes of convenient identification and understanding, and that some members of those classes may appear to be more 'central' members than others. For example, a penguin makes a worse prototypical bird than a robin or a blackbird, because it can't fly and is an odd shape. The purpose of a prototype and the human ability to group things around it is basically because things in the real world differ from one another, but that some things (birds, chairs, cars, democracies) share enough common features for us to be able to identify them as instances of the 'same thing':

> 'The world consists of a virtually infinite number of discriminably different stimuli. One of the most basic functions of all organisms is the cutting up of the environment into classifications by which non-identical stimuli can be treated as equivalent.' (Rosch *et al*, 1976:383).

So peripheral members of a group are open to classification as part of more than one such group (for example a table lamp is a peripheral member of the categories 'furniture' and 'electrical household appliance'). In terms of document genres, then, we might think that there are forms that are 'formier' than others, and newspapers that are more newspapery. By extension, there are elements of such documents – pattern solutions – that make more or less prototypical solutions to their problems.

For example, Figure 3 shows quite a good solution to the problem of eliciting a name on a form and is typical of the forms genre in its current state in the UK. Users of the form in Figure 4, however, often fail to supply the name correctly, because the sentence-completion solution used to elicit the name is now largely obsolete. The same solution seems quite at home, however, in the children's party invitation, which is a more peripheral member of the forms genre.



YOU

1.1 Title, enter MR, MRS, MISS, MS, or other title

□□□□□□

Surname

□□□□□□□□□□□□□□

1.2 First name(s)

□□□□□□□□□□□□□□

□□□□□□□□□□□□□□

HM Revenue & Customs

Potential repayment to the estate

Please use these if you write or call. It will help to avoid delay.

Your reference

Our reference /

National Insurance number

If you need more help or advice call

Name

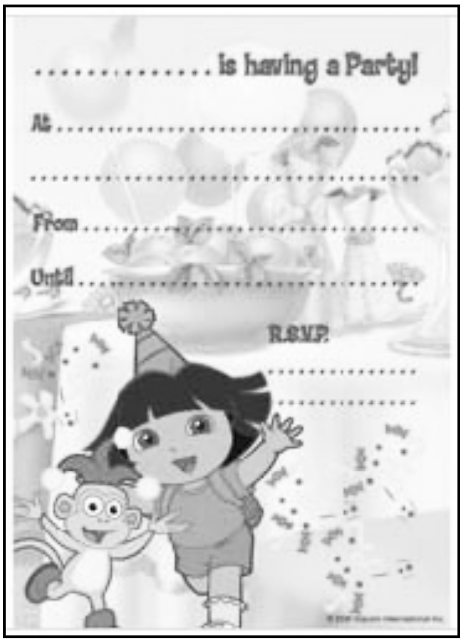
Phone

Office address/date stamp

Estate of the late ____________________

I have sent you this form so that I can finalise the tax position up to the date of death. If you are dealing with the deceased's estate please complete this form. If you are not, please give it to the person who is. They will normally be called the Personal Representative or Executor and may be a solicitor or accountant. **It is important that this form is completed and signed personally by the Personal Representative** and not by an agent or professional adviser acting on their behalf.

Figure 3: Two ways to elicit someone's name. The left-hand example, using the CHARACTER BOXES *solution is more prototypical of the current state of the forms genre. The right-hand example uses the solution* SENTENCE COMPLETION *which is largely obsolete, and therefore peripheral.*



…………… is having a Party!

At …………………………………

…………………………………

From …………………………………

Until …………………………………

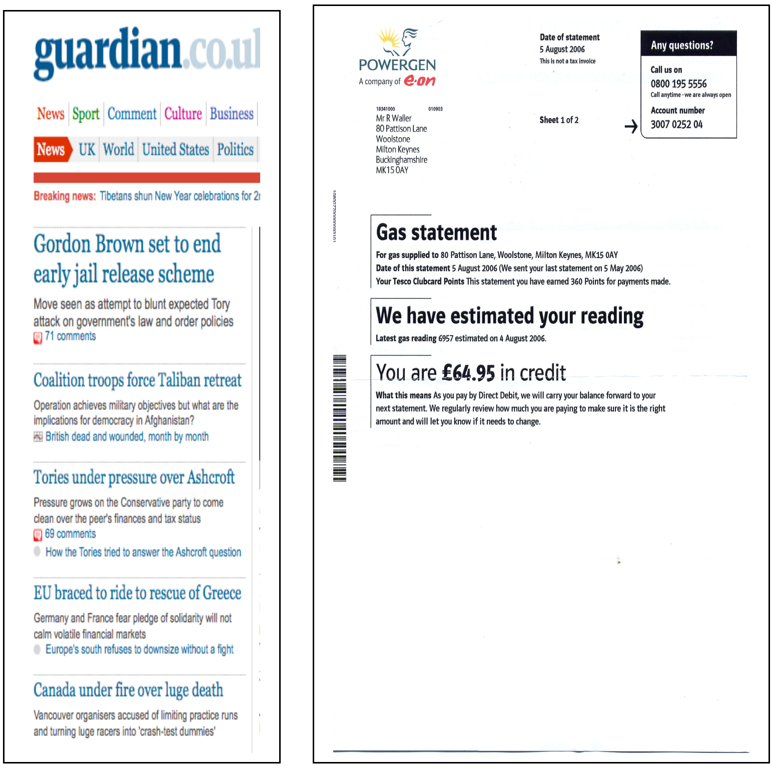
R.S.V.P.

……………

……………

Figure 4: The SENTENCE COMPLETION *solution seems quite at home in this prototypical party invitation (making it also a peripheral member of the forms genre).*

While patterns can occur within different genres, it may well be the case that many of them have a 'home genre' in which they are an essential feature. For example, the pattern LIST OF INGREDIENTS is an essential feature in its home genre 'recipe book', but it also occurs in the genre 'form' (where users might be given lists of key information to gather before starting).



guardian.co.uk

News | Sport | Comment | Culture | Business

News | UK | World | United States | Politics

Breaking news: Tibetans shun New Year celebrations for 2

Gordon Brown set to end early jail release scheme

Move seen as attempt to blunt expected Tory attack on government's law and order policies

71 comments

Coalition troops force Taliban retreat

Operation achieves military objectives but what are the implications for democracy in Afghanistan?

British dead and wounded, month by month

Tories under pressure over Ashcroft

Pressure grows on the Conservative party to come clean over the peer's finances and tax status

69 comments

How the Tories tried to answer the Ashcroft question

EU braced to ride to rescue of Greece

Germany and France fear pledge of solidarity will not calm volatile financial markets

Europe's south refuses to downsize without a fight

Canada under fire over luge death

Vancouver organisers accused of limiting practice runs and turning luge racers into 'crash-test dummies'

POWERGEN
A company of e·on

Date of statement
5 August 2006
This is not a tax invoice

Any questions?
Call us on
0800 195 5556
Call anytime - we are always open
Account number
3007 0252 04

18341000 010903
Mr R Waller
80 Pattison Lane
Woolstone
Milton Keynes
Buckinghamshire
MK15 0AY

Sheet 1 of 2

Gas statement

For gas supplied to 80 Pattison Lane, Woolstone, Milton Keynes, MK15 0AY
Date of this statement 5 August 2006 (We sent your last statement on 5 May 2006)
Your Tesco Clubcard Points This statement you have earned 360 Points for payments made.

We have estimated your reading

Latest gas reading 6957 estimated on 4 August 2006.

You are **£64.95** in credit

What this means As you pay by Direct Debit, we will carry your balance forward to your next statement. We regularly review how much you are paying to make sure it is the right amount and will let you know if it needs to change.

Figure 5: The NEWS HEADLINES *pattern in its home genre (left) and in a gas bill (right)*

Figure 5 shows how the pattern NEWS HEADLINES has been transported from its home genre, ‘newspaper’ to the genre ‘gas bill’. The resulting bill thus departs from its genre, but it nevertheless works because the headlines enable a effective reading strategy. Indeed, the use of headlines is an implicit intertextual reference to the newspaper genre, suggesting to readers that the reading strategy they use there (that is, a quick preview possibly, but not necessarily, followed by a detailed read of stories that interest them) is also appropriate for reading a bill. In time, if successful enough to imitate, the energy bill genre may shift.

While we are looking for a way to identify possible members of a set of solutions in a given pattern, therefore, we should note the following:

- Available solutions may be constrained by genre, but are also judged on their functionality in context, and the quality of their execution.
- Within the set created by the genre constraint, members will be more or less prototypical.

A hypothesis might be that solutions that are less prototypical might (a) be harder for users to identify visually as belonging to the pattern or the genre, and might therefore cause slower response rates and/or higher error rates, and (b) might, if they are less constrained, be more likely to turn up as possible solutions to other patterns. In this case, the more prototypical a solution is to pattern B, the more likely it is to cause confusion when used as a solution to pattern A – even if it appears within A’s set of reasonable possible solutions.

Cohen and Snowden (2008) have indeed demonstrated a correlation between the familiarity of document elements to readers and their performance in literacy tests. They use the term ‘document mental model’ to describe the kind of genre-specific knowledge required by competent readers that should be anticipated by competent document designers.

> ‘Readers are likely have a different mental model for each specific document type with which they are familiar. When confronted with a document, readers may recall and use these mental models, which, if accurate, should aid them in locating the vital information. For example, menus often contain the price of a dish to the right of the listing for that dish. For those with an accurate “menu” mental model, a request to locate price should be facilitated when the information is near the predicted location and inhibited when it is not.’ (page 19).

7 The research context of this discussion

The pattern language approach introduced in this paper is part of a wider research programme that includes, firstly, the building of a document corpus, so we can demonstrate the frequency of patterns within a particular domain (in the first instance, financial services documents), and, secondly, the testing of documents (selected to include patterns and genres of greater or lesser prototypicality, as well as other variables such as the strength of graphic and linguistic signalling) with users who come with different levels of experience and financial capability.

References

ALEXANDER, C., ISHIKAWA, S., & SILVERSTEIN, M. (1977). *A pattern language. Towns. Buildings. Construction.* New York: Oxford University Press.

ALEXANDER, C. (1979). *The timeless way of building*. New York: Oxford University Press.

BATEMAN, J. (2008). *Multimodality and genre: a foundation for the systematic analysis of multimodal documents*. London: Palgrave Macmillan.

COHEN, D. J., & SNOWDEN, J. L. (2008). The relations between document familiarity, frequency, and prevalence and document literacy performance among adult readers. *Reading Research Quarterly, 43*(1), 9-26.

CROFTS, K. (2009) *Companion report: Patterns and Problems*. Unpublished project report for MA in Information Design, Department of Typography and Graphic Communication, University of Reading, UK.

DELIN, J., BATEMAN, J., & ALLEN, P. (2002). A model of genre in document layout. *Information Design Journal, 11*(1), 54-66.

GAMMA, E., HELM, R., JOHNSON, R., & VLISSIDES, J. M. (1994). *Design patterns: elements of reusable object-oriented software*. Addison-Wesley.

HOEY, M. (1983). *On the surface of discourse* London: Allen & Unwin.

HUNSTON, S., & FRANCIS, G. (2000). *Pattern grammar: a corpus-driven approach to the lexical grammar of English.* Amsterdam: John Benjamins.

ROSCH, E. H. (1973). Natural categories. *Cognitive Psychology, 4*, 328-350.

ROSCH, E. H., MERVIS, B., GRAY, W., JOHNSON, D., & BAYES-BRAEM, P. (1976). Basic objects in natural categories. *Cognitive Psychology 8*, 382-439.

SCHÖN, D. A. (1983). The reflective practitioner: how professionals think in action. London: Temple Smith.

SCOTT, BILL (2006) contribution to a web discussion on pattern libraries: *Design Patterns: Part 3* (http://www.lukew.com/ff/entry.asp?350)

TAYLOR, J. R. (2003). *Linguistic categorization*. Oxford University Press.

TIDWELL, J. (1997). *Common ground: a pattern language for human-computer interface design*, from http://www.mit.edu/~jtidwell/common_ground.html

TIDWELL, J. (2005). *Designing interfaces: patterns for effective interaction design*. O'Reilly.

TWYMAN, M. (1979) A schema for the study of graphic language. In *The Processing of Visible Language*, (Eds, Kolers, P.A., Wrolstad, M.E. & Bouma, H.) Plenum, New York

WALLER, ROBERT (1987). *The typographic contribution to language: towards a model of typographic genres and their underlying structures.* Unpublished PhD thesis, Department of Typography and Graphic Communication, University of Reading, August 1987.

WALLER, ROBERT (1990) ‘Typography and discourse’, in R Barr et al (eds), *Handbook of Reading Research*, vol II, New York: Longman, 341–380.

WITTGENSTEIN, L. (1953). *Philosophical investigations*. Oxford: Basil Blackwell.

Spatial Coding and information retrieval in multimodal documents

Claudio Vandi (1), Thierry Baccino (2)

(1) Laboratoire CHArt, Université de Paris 8 - LUTIN UMS CNRS 2809
vandi@lutin-userlab.fr
(2) LUTIN UMS-CNRS 2809
baccino@lutin-userlab.fr

Abstract In this paper we study the effects of Spatial Coding (Murray, Kennedy, 1987; Kennedy, 1992) on the way readers process semantic information while searching for a target in multimodal (text- image) documents. Our aim is to understand how spatial indexes are used by readers in information retrieval tasks, and what are the interactions between semantic distance and spatial distance. We present an eye tracking experiment in which 24 participants were asked to search for information in simple multimodal digital documents while we manipulated their ability to associate spatial indexes to semantic content and changed the location of the target text between the inspection and the search phases. The results show that participants' performances worsened with the increasing of the semantic distance between the target text and the text that took its place. Results also show that readers can rapidly adapt their information searching habits to the spatial organization of the document.

Keywords: spatial coding, multimodal, information searching, reading, eye tracking, habits.

1 Introduction

1.1 Spatial organization and semantic processing in texts

Spatial organisation can be a powerful tool for signalling text organisation and is likely to have a direct influence on information searching skills and text comprehension abilities of readers. Previous studies on reading and information retrieval put forward a hypothesis referred to as the "spatial coding hypothesis" (Kennedy, 1992; Kennedy et al., 2003), according to which "readers must maintain, and use, a level of representation of text that involves the computation of spatial coordinates" (Kennedy, Murray, 1987). Authors who support this hypothesis show, for example, that readers are capable of localising the first word of an anaphora by making a single and very precise saccade to its position, thus presupposing that they first *coded* this position, even if they weren't explicitly asked to do so (Kennedy, 1982; Murray, Kennedy, 1987). Another study shows the ability to code spatial position of words is a fundamental step in becoming an expert reader (Murray, Kennedy, 1988). Other studies show that readers that are able to associate semantic and spatial information have a better understanding of texts, compared to subjects that receive only non localised information (Virbel et al., 2005). This hypothesis is also consistent with "position special" theories of vision that claim that spatial features (the "where") play a special role in vision, compared to other visual features like colour and shape (the "what") (Van der Heijden, 1993; Van der Heijden et al., 1999; Spivey, 2001). Furthermore, the "spatial coding" phenomenon is also considered as one of the proof for the validity of the "extended mind" theory (Clark, Chalmers, 1998), according to which readers do not need to create a mental copy of the whole document while reading it, but just place some spatial indexes on the page in order to use it as an "external memory" that can be accessed to retrieve information when needed (O'Reagan, 1992; Richardson, Spivey, 2000).

1.2 Spatial coding and its relation to semantics in multimodal documents

Even if it is hard to find someone who denies that people retain *some kind* of spatial information while reading, to understand the role that spatial indexes can have on text signalling organisation and text comprehension we still need to answer a question concerning the *target* of this spatial coding. The main controversy in this respect concerns the role of semantic information, and can be expressed as follows: is spatial coding associated with the position of graphical elements (i.e. memory of the word's shape) or is it associated with the position of meaningful entities (i.e. memory of the semantic entity) ? For example in the following example from Kennedy and Murray (1987):

(1) The novels in the library had started to go mouldy with the damp. novels

When readers go back to the anaphora "novel" is their search aimed at the *meaning* of the word or do they just remember its *aspect* ? Our hypothesis (that we share with Baccino, Pynte, 1994) is that *spatial coding is semantically motivated* and consists in associating semantic labels with locations in space, so that actions (be they real actions or oculomotor ones) that are addressed to these locations will be semantically charged. This coding allows readers to develop dispositions towards action: having attributed a semantic value to specific locations in space, they know where to look for information and don't have to scan the entire scene before acting.

As far as the organisation of multimodal documents is concerned, our hypothesis about spatial coding can be articulated in three more detailed ones:

1. If spatial coding is based on spatial localisation of *semantic* information (i.e. not on its surface aspect) we should be able to observe its effects even if the cue and target do not share the same surface aspect (e.g. they belong to different modalities).

2. If spatial coding consists in attributing semantic labels to locations in space, we expect that if the semantic content of the text changes after the reader has labelled it, when he

will search for information again, he will behave in different ways if the new content is semantically associated or not associated to what he expects to find.

3. If readers rely on spatial coding for semantically organising a text, we expect that even when they are presented with documents, the spatial organisation of which is unconventional, they will adapt their research habits to the spatial structure of these documents.

Verifying these hypothesis would allow us to better understand how spatial and semantic relations influence the cognitive organisation of a text and how spatial and semantic information should be coupled if we want to create coherent and easy-to-search multimodal documents. This would be particularly important for the design of digital documents (web pages, digital documents for education) where document organisation changes in a dynamic way according to the support being used (desktop monitor, mobile device) or the contents being displayed (e.g. a news web page). To verify our hypothesis we performed an eye tracking experiment in which participants were asked to search for information in a series of simple multimodal digital documents composed of three texts and one image.

2 Methods

2.1 Participants and material

24 participants took part in our eye tracking experiment. They were divided into two equal groups: a *test group* and a *control group*. Both groups were presented with a series of 24 trials on a screen. For the test group, each trial was composed of two documents: the first one made up of three texts organised in a circle (left, top, right) and the second of an image that represented a common[1] object or animal in the middle of the screen surrounded by the three texts. For the control group, the first document presented only the image, the second document was composed of an image and three texts, identical to those of the test group. Each text was 320 characters long on average and described a common object or animal. In each document, only one of the three texts was strongly associated with the image. For the other two texts, one was only weakly associated with the image, and the other was non-associated. The intensity of this semantic association was computed using *Latent Semantic Analysis* (LSA) *one to many* comparison (Landauer et al., 1998; Bellissens et al., 2004).

2.2 Procedure

For the test group, in each trial, the documents were presented in two phases: an inspection phase and a search phase. During the inspection phase, participants had 10 seconds to skim the document. After ten seconds, a black screen appeared for 1 second and then the search phase began. During the search phase, the participants had to find the text that was strongly associated with the image. Participants were asked to locate this target text and press the spacebar. To make sure they located the target, they were asked to fix the correct text in their gaze while pressing the spacebar. Once they pressed it, they passed on to the next document.

[1] Between 10 et 120 occurrences in the « Lexique » database http://www.lexique.org/listes

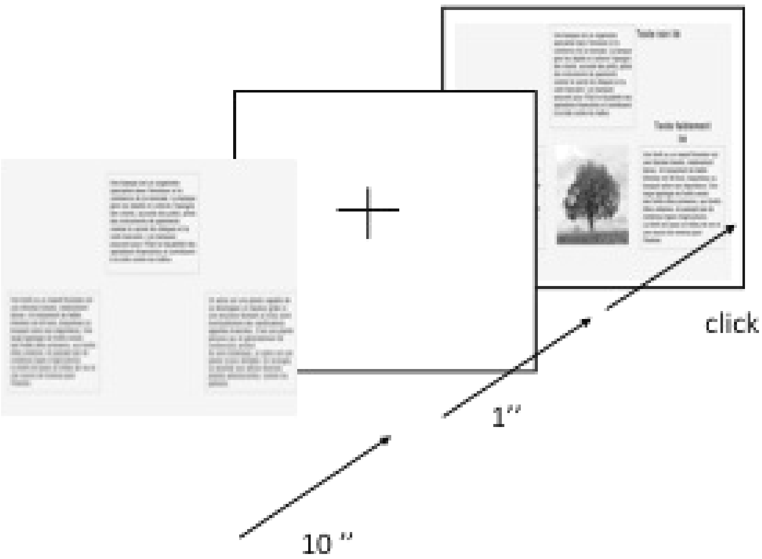


Figure 1 Schematic stimulus display for the test group. An inspection phase (10 seconds), followed by a neutral screen with a central cross (1 second) and a search phase (until the participant press the spacebar).

Between the inspection and the search phase, the content of the texts was the same. As far as their positions are concerned, in 50% of the cases it didn't change (a "No change" condition); in 25% of the cases, the target text swapped its position with the *weakly* associated test (a "Weak change" condition); in another 25% of the cases, it swapped its position with the *non-associated* text (a "Non-associated change" condition). The initial positions of texts were counterbalanced so that probability of the associated text appearing in any of the three positions was the same.

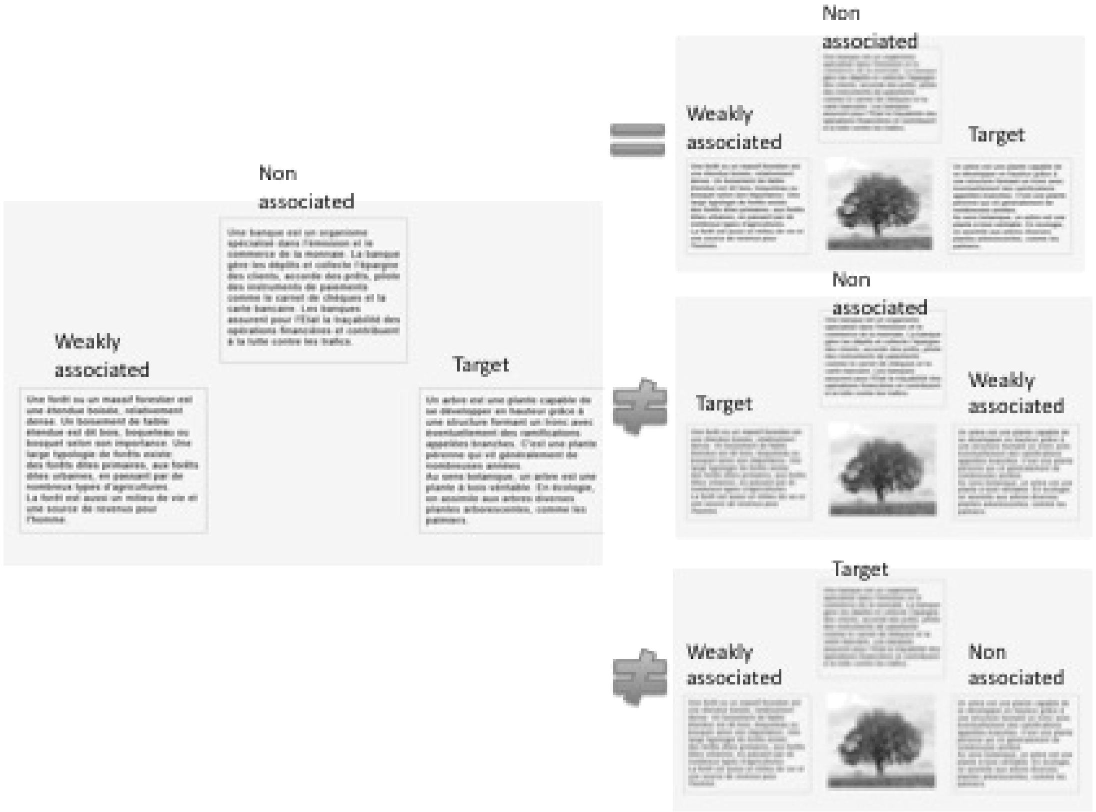


Figure 2 Schematic example of a trial for the test group. Left image : inspection phase. Right images : possible alternatives for the search phase. Target text: tree. Weakly associated text: forest. Non associated text: bank.

The procedure was the same for the control group, except that during the inspection phase, participants only saw the image, without texts. In this way, they didn't receive any priming about the spatial position of information during the inspection phase. During the search phase, the screens presented were identical to those of the test group. The task was also the same: during the inspection phase, the participants had to locate the associated text and press the spacebar. But since these participants had never seen the texts before the search phase, they could not experience any sort of spatial modification with respect to the inspection phase in which they saw just the image.

2.3 Eye-Movement Recording

Eye movements were recorded by means of a Tobii 1750 eye tracker. Response Times were recorded too. To record scanpaths, the screen was divided into four Areas Of Interest (AOIs) one for the central image, and three for each of the text zones (left, top, right).

3 Data Analysis

3.1 Error rate

For the scope of this study, an error was defined as each trial in which a subject pressed the spacebar while looking at the wrong text. The error rate (averaged over all participants) was

10% with no significant difference for the three conditions (no change, weak change and non related change) within the test group (t(22)= -0.85 ; p =.96*) and no significant differences between the test and the control group (F(2,22)= 1.11 ; p = .34*).

3.2 Eye-Movement Data

We analysed Eye-Movement and Response Time only for error-free trials. Our data shows that Response Times, overall fixations duration and fixations number are highly correlated (RT-FixNumber r = 0.97 ; RT-FixDuration r= 0.97; FixNumber-FixDuration r= 0.95). For the sake of simplicity and readability we have chosen to present here only the data about Overall Fixations Duration and Scanpaths. For scanpaths, we analysed the number of AOIs fixed by the subjects and the first fixation location. Since all the first fixations were located on the central image we decided to analyse the first fixation outside the central image AOI.

Overall fixations duration (ms ± SD)			Number of fixed AOIs (number ± SD)	
Condition	Test group	Control group	Test group	Control group
No change	1407.5 ± 370.7	2801 ± 1042.22	2.72 ± 0.34	4.03 ± 0.67
Weak change	1843 ± 904.4		3.26 ± 0.74	
Non-associated change	2058.7 ± 586.3		3.24 ± 0.31	
Average	1637.16 ± 343.3		3.08 ± 00.33	

Table 1: Overall fixations duration and numer of fixed AOIs.

First fixations location (% ± SD)			
Location	Test group inspection phase	Test group search phase	Control group search phase
Left	57 % ± 0.03	43% ± 0.16	65% ± 0.26
Centre	41 % ± 0.03	26% ± 0.18	15% ± 0.19
Right	4% ± 0.01	30% ± 20	18% ± 0.24

Table 2: First fixations location

4 Results

4.1 Information searching performances: effects of spatial coding

To verify the effects of receiving spatial information in the *inspection* phase on information searching performances in the *search* phase, we compared Eye Movements of the test group (the spatial coding group) with those of the control group (no spatial information received). Our results showed that the performance of the test group were significantly better than those of the control group. The overall duration of fixation was shorter for the test group *G1 (t(22)= -3.67 ; p < .01*). As far as the participants' scanpaths are concerned, the results showed that the test group fixed a smaller number of AOIs before locating the target text (*t(22)= -4.35 ; p < .001*). In fact, participants in the control group browsed through an average of 4 textual AOIs before locating the target. This means that most of the time they read through all three texts (3 textual AOIs fixed) and then re-fixed their gaze on the chosen one (4th fixation). The percentage of trials in which the subject made at least one re-fixation is significantly greater for the control group: 45% versus 10% (*G1 (t(22)= -3.40 ; p < .01*).

4.2 Information searching performances : effects of spatial and semantic distance

As far as the effects of the semantic distance on information searching performance are concerned, the results for the test group showed that information searching performances were significantly better when spatial organisation of texts didn't change (the "No change" condition) compared to when the location of the target text changed (the "Weak change" and "Non-associated change" conditions). As showed in table 1, under the "Weak change" and "Non-associated change" conditions, the number of fixations was significantly greater (*t(12)=*2.7 ; p < .05) as was the number of fixed textual AOIs (t(12) = 3.76 ; p < .01). Looking at the results in greater detail, we observe that those obtained under the "Weak change" condition (when the target text changes place with the weakly associated text) were significantly different from those obtained under the "Non-associated change" condition (when the target text changes position with the Non-associated text). In particular, pair-wise comparisons showed that overall fixation duration was significantly longer under the "Non-associated change" condition than under the "Weak change" condition (F (1,10) = 9.32 ; p < . 05), which means that the subjects had lower performances when the Non-associated text replaced the target text.

Concerning the participants' scanpaths, pair-wise comparisons showed that when the target text kept its original position (the "No change" condition) the average number of AOIs that the participants needed to fix with their gaze before locating the target was significantly smaller (F(1,10) =17.0 ; p < .005). The results presented in table 1 also showed that when the target text didn't move, the performance of the test group that had already read the texts in the inspection phase was far better than that of the control group, which had not read the texts before the search phase. Under the "Weak change" conditions, the performances of the test group approach those of the control group, and get even closer (i.e. worse) under the "Non-associated change" condition.

4.3 Information searching habits: effects of spatial cue

To conclude, we analysed first fixation location in order to detect the participants' habits (the point from which they begin) when performing an information searching task on a multimodal document. These results are presented in table 2. They show that the participants of the test group (during the inspection phase) and participants of the control group (during the search phase) tended to begin their information searching activity from the left text, following the well-established occidental left-right reading habit (57% and 65% for the left, 41% and 15% for the centre, 4% and 18% for the right). For the test group, pair-wise comparisons showed that, during the inspection phase, the left position was fixed with the subject's gaze significantly more often compared to the situation in the other positions *(F(1,11)= 18.6 ; p < 01*). On the other hand, during the search phase , the test group shifted to a more balanced distribution (43% left, 26% centre, 30% right) with no significant difference *(F(2,22)= 1.78 ; p = .19 **) once they had explored the spatial distribution of texts during the inspection phase.

5 Discussion

The aim of this paper was twofold: i) To determine the relations between spatial coding and semantic processing. In particular we wanted to understand if semantic distance can have an effect on spatial processing and if this can also be observed in those situations in which cue and target belong to different modalities. ii) To prove that readers are able to modulate their reading habits in order to adapt their research strategies quickly to the spatial organisation of text. The results we obtained support our hypothesis.

Results presented in 4.1 show that spatial coding effects on reading and information searching can indeed be observed even in documents in which the target and the cue belong to two different modalities (e.g. textual target cued by an image). We observed that participants who could rely on spatial coding (the test group) and spatially uncued participants (the control group) each had different searching performances and scanpaths. Furthermore, we observed that the participants in the control group needed to skim all the texts and then re-fix their gaze on the right one before answering. According to our hypothesis, this result shows not only the participants had an advantage when they could attribute a spatial identity to each semantic content, but also that when they lacked this kind of spatial information (the control group), they needed to identify each text spatially before deciding on its semantics.

The results presented in 4.2 show that semantic distance has indeed an effect on spatial coding. As expected, when the target text was replaced by weak or non-associated texts, the participant's performances worsened with the increasing of the semantic distance between the target text and the text that took its place: participants had better performances (shorter overall fixations time, fewer AOIs fixed before answering) under the Weak-change condition compared to under Non-associated change condition. We interpret these results as being caused by the co-activation of spatial and semantic information during the search activity. This activation makes it easier for the reader to process weakly associated texts because they are close to his expectations. These results are consistent with previous findings concerning co-activation of spatial and semantic information during information searching (Baccino, Pynte, 1994).

The results presented in 4.3 show that readers can easily change their reading habits and adapt their oculomotor behaviour to the local structure of information. In particular, participants in the test group easily broke the "begin from the left" rule during the search phase, when they understood that relevant information could appear with equal probability in any of the three locations. These results are consistent with previous studies about standard for web interfaces that show that users can easily adapt to unconventional structures if these structures are coherent and semantically motivated (McCarthy et al., 2003).

To resume, our experiment showed that spatial coding effects are observed even in multimodal documents, that spatial organisation is strongly intertwined with the text's semantic organisation, and that readers easily adapt their research habits to the way information is spatially and semantically organised in a document. As far as the signalling of text organisation is concerned, this study shows that spatial organisation can be a powerful tool for signalling the semantic organisation of a document. Spatial distance can be use to signal a semantic distance and to attribute a semantic value to a location in space. Once they know were a semantic information is spatially located readers can adapt their information searching habits to the document's organisation and find information more easily. To conclude, this study allows us to make at least three recommendations to design multimodal documents such as web pages or digital documents for education effectively: *i.* Clearly link each item of content with a spatial location to obtain documents in which information can be easily searched and retrieved *ii.* Maximise the coherence between *semantic* and *spatial* distance to signal text organisation effectively *iii.* Shape readers' reading habits by adopting a coherent spatial-semantic architecture on which readers can rely when they search for information.

References

BACCINO T., PYNTE J. (1994). Spatial coding and discourse models during text reading. *Language and cognitive processes 9-2*, 143- 155.

BELLISSENS C., THÉROUANNE P., DENHIÈRE G. (2004). Deux modèles vectoriels de la mémoire sémantique: description, validation et perspective. *Le Langage et L'Homme* 39*,* 101-122.

CLARK A., CHALMERS D. (1998). The extended mind. *Analysis* 58, 10-23.

KENNEDY A. (1982). Eye Movements and Spatial Coding in Reading. *Psychological Research* 44, 313-322.

KENNEDY A. (1992). The Spatial Coding Hypothesis. In K. Rayner (Ed.) *Eye Movements and Visual Cognition*. New York: Springer-Verlag (379 - 397).

KENNEDY A., MURRAY W.S. (1987). Spatial coding and reading: Some comments on Monk (1985). *Quarterly Journal of Experimental Psychology* 39A, 649-718.

KENNEDY A., BROOKS R., FLYNN L., PROPHET C. (2003). The reader spatial code. In J. Hyönä, R. Radach, H. Deubel (Ed.) *The mind's eye: Cognitive and applied aspects of eye movement research*. Amsterdam : North-Holland. (193–212).

LANDAUER T.K., FOLTZ P.W., LAHAM D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes* 25, 259-284.

McCarthy J.D., Sasse A.M., Riegelsberger J. (2004). Could I have the menu please? An eye tracking study of design conventions. Proceedings of *People and Computers XVII*, 401-414.

Richardson D.C., Spivey M.J.(2000). Representation, space and Hollywood Squares: looking at things that aren't there anymore. *Cognition* 76, 269-295.

Spivey M.J., Geng J.J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological Research* 65, 235-241.

Van der Heijden A.H.C. (1993). The role of position in object selection in vision. *Psychological Research* 56, 44-58.

Van der Heijden A.H.C, Musseler J., Bridgeman B. (1999). On the Perception of Position. *Advances in Pshychology* 129, 19-37.

Virbel, J. Schmid, S., Carrio, L., Dominguez, C., Woodley, MPP, Jacquemin, C., Mojahid, M., Baccino, T., Garcia-Debanc, C. (2006). Approches cognitives de la spatialisation du langage: le cas de l'énumération. In. C. Thinus-Blanc, J. Bullier (Eds.) *Agir dans l'espace,* Paris Maison des Sciences de l'Homme (233-254).

Text signals in the aircraft maintenance documentation

Herimanana ZAFIHARIMALALA, André TRICOT
In collaboration with EADS IW

Université de Toulouse2 – Le Mirail
herimanana.zafiharimalala@univ-tlse2.fr,
andre.tricot@tououse.iufm.fr

Abstract: After a brief description of the aircraft maintenance documentation trough its functions and content, this paper presents results from two types of studies about the use of text signals in the maintenance documentation used in the aeronautical field. The first study is based on the observations and interviews of the maintenance operators during their task; the second study concerns ergonomic inspection of the documentation based on ergonomic criteria about structure and content. The two results confirm the problem linked to the information presentation (characters size, bad quality of figures, etc.) and structure and allow also concluding that aircraft maintenance documentation must be improved by the use of more efficient text signals in order to make easier the use the documentation at the work place and to avoid maintenance errors.

Keywords: Aircraft maintenance documentation, procedural documentation, text signals.

1 Introduction

Maintenance and inspection are the major factors of aircraft accident for 12 % (Hobbs, 2000) because of errors made during the operation. According to Chaparro et al. (2002) and the FAA (Federal Aviation Administration), the cause of most of the maintenance errors is the documentation used to guide maintenance tasks. For Lattanzio et al. (2008), procedural errors in aeronautic maintenance have different forms and are due to both document quality and users errors (see Figure 1).

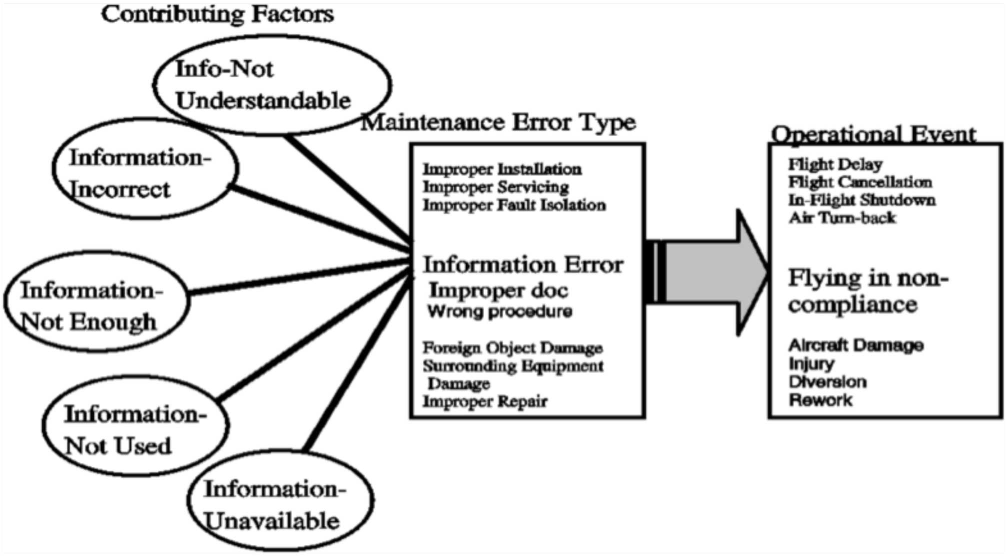


Figure 1 The relationship between information quality (contributing factors), maintenance error types and operational events (Lantazio et al., 2008).

Consequently, the usefulness (relevance of the content) and the usability (relevance of the access to the content) of the maintenance documentation must be improved in order to encourage at least the operators to use the documentation. Indeed, according to Chaparro et al. (2004), the 64% of maintenance technicians report using their own manner to execute maintenance procedure. Some routine maintenance tasks are completed in a different manner than the procedure in the documentation (McDonald et al., 2000). Van Avermaete and Hakkeling-Mesland (2001) also report that 34 % maintenance technicians turn away from the documentation.

In view of the issue of a documentation which is not used systematically, or if used constitutes a factor of human errors, this article concerns information presentation in the maintenance documentation in order to make an inventory of the features (signals) used for this purpose, their role and to determine the points to be improved. Indeed, according to Caro et Bétrancourt (2001), beyond the document content, the structure and the layout have a repercussion in the reader’s representation, and consequently in his later task performance. This influence is particularly important in the technical documentation because of the disastrous human and equipment consequences of the bad interpretation of the content.

The aircraft maintenance documentation will first be described by its different functions and its content. Secondly, ergonomic inspection results about the signals in the aircraft maintenance documentation will be presented. An inventory of signals rarely or not used in the documentation will then be established. Finally, information from a field study about the role of signals in the aircraft maintenance documentation use will be presented.

2 Brief description of the aircraft maintenance documentation

2.1 The functions of the aircraft maintenance documentation

The maintenance documentation used in aeronautic domain has three functions: a support of maintenance task, a legal document and a support of training.

2.1.1 The documentation as a support of the maintenance task

The technical documentation used in the aeronautical maintenance is a procedural document: it is used to complete an action, which is the maintenance task. It aims to guide the maintenance operator trough a list of instructions in the task execution (Montmollin, 1997; Cellier, 2005). The documentation corresponds to the prescribed task of the operator: it describes how the work must be completed. The documentation consultation is a part of the maintenance task. It is strongly linked to the operator's task.

2.1.2 The documentation as a legal document

In the aeronautic field, documentation use is legally obligatory. It constitutes a proof that maintenance operation has been executed in accordance with the instruction. Thereby, the maintenance documentation is a mean certifying safety and security of the aircraft system after the maintenance operation. Indeed, the maintenance operator must sign in the documentation used at the end of the maintenance operation. The signature has an important legal role. For the maintenance center, it constitutes a protection in case of a conflict with the customer after the delivery of the aircraft.

2.1.3 The documentation as a support for the training

The maintenance documentation has also a role of a support for the training of maintenance operators. In developing countries, as maintenance operators have a thorough knowledge of the job since their training, the maintenance documentation is mainly used as a support to the maintenance task, contrary to the least developing countries where users (operators) most frequently discover the documentation during the task completion.

2.2 The maintenance documentation content

Aircraft maintenance documentation makes available different information such as maintenance tasks references and titles, a list of figures (a figure illustrates the procedure of the task selected by the user), the identification code of the task (called *code FIN*), warnings message, caution message, the *job set-up information* which corresponds to the information about the task context and material required (e.g. Fixtures, Tools, Test and Support Equipment) with the reference, the quantity and the designation ; the work zones (job location) and the access panels (i.e. how to access to the work zone) ; the *Job set-up* which gives details about other information required before the completion of the main procedure, such as the *safety precautions (e.g. "Put the warning notice in the cockpit to tell persons not* *to operate the landing gear”),* a check to do before the main procedure, information about the aircraft maintenance configuration (e.g. “*make sure that the PARK BRK control switch is set on ON*”) ; the *Procedure* (the main procedure which constitutes the object of the maintenance operation). A procedure can be a set of subtasks. Thus, the *Procedure* section contains information about the reference, the object and the procedure of each subtask and sometimes a caution; the *Close-up* which consists in the information about the tasks to be completed after the maintenance operation (e.g. setting right the aircraft, cleaning the workplace, etc.).

The description of the task in the maintenance documentation corresponds to the MAD formalism (Analytic Description Method) described by Sébillotte (1991). According to the MAD formalism, task disposes of different characteristics:

- An identification: title, number (reference)

- Elements: the aim of the task, an initial state of the task, pre conditions (constraints concerning the initial state of the object, which corresponds to the warnings in the maintenance documentation), the “body” of the task (information about how the task is executed, i.e. the subtasks), the postconditions (constraints concerning the final state of the object, which corresponds to the close-up part), the final state of the object (a list of the elements modified by the task)

- The attributes: specific characteristics of some subtasks (optional, iterative, priority, interruptible). According to the aircraft maintenance documentation, some tasks cannot be executed under some weather conditions or need specific qualification.

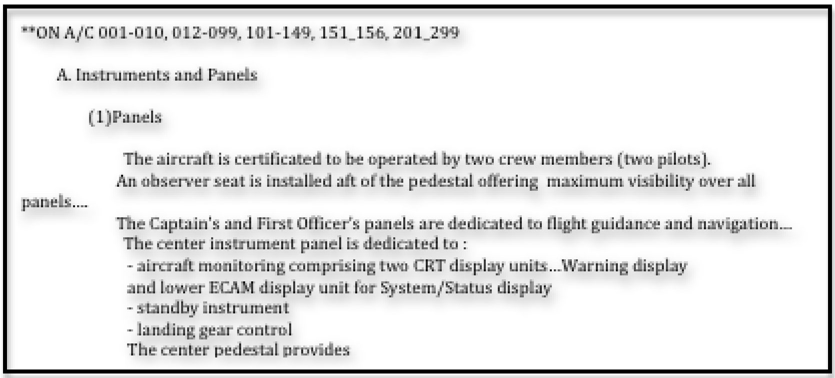
3 Signals in the aircraft maintenance documentation

The inventory of the following signals means is based on the study of Caro and Bétrancourt (2001) about the ergonomic study of electronic document, and on the definition given by Terrier et al. (2005) of features used to draw reader’s attention. For Terrier et al. (2005), important elements and information within a text are distinguished by the use of different types of means (called MFM or Mise en Forme Matérielle): typographical properties (bold characters, italic characters, etc.), layout properties (space, indentation), and colour. The lack of those means makes reading difficult and cognitively costly because of the necessity to read the text entirely to find the information required.

3.1 Layout properties

- Space: in some maintenance documentation, there is no space between items.

Text signals in the aircraft maintenance documentation



**ON A/C 001-010, 012-099, 101-149, 151_156, 201_299

A. Instruments and Panels

(1)Panels

The aircraft is certificated to be operated by two crew members (two pilots).
An observer seat is installed aft of the pedestal offering maximum visibility over all panels....

The Captain's and First Officer's panels are dedicated to flight guidance and navigation...
The center instrument panel is dedicated to :
- aircraft monitoring comprising two CRT display units...Warning display and lower ECAM display unit for System/Status display
- standby instrument
- landing gear control
The center pedestal provides

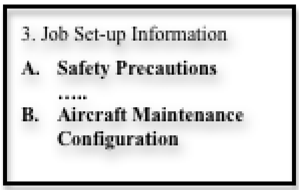
Figure 2 Text spacing in the maintenance documentation

The lack of space between different items can induce difficulty to the user when reading the documentation. According to Caro and Bétrancourt (2001), more space improves the reading speed.

- Listing: numerical and alphabetical listings are used in the aircraft maintenance documentation.

The numerical listing can be a number followed by a dot to indicate a section in the procedure (e.g. 4. Job set-up) or a number between brackets to indicate a step in the procedure (e.g. (1) Put the safety barriers in position).

The alphabetical listing can be a capital letter followed by a dot used for the subsection or a minuscule letter between brackets to list the different actions of a step in a procedure



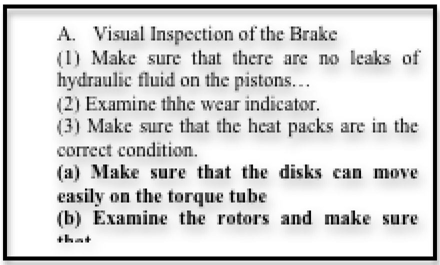
3. Job Set-up Information

A. Safety Precautions

.....

B. Aircraft Maintenance Configuration

Figure 3 *alphabetic listing using Capital letter + dot*



A. Visual Inspection of the Brake
(1) Make sure that there are no leaks of hydraulic fluid on the pistons…
(2) Examine thhe wear indicator.
(3) Make sure that the heat packs are in the correct condition.
(a) Make sure that the disks can move easily on the torque tube
(b) Examine the rotors and make sure that

Figure 4 *alphabetic listing using minuscule letter between brackets*

3.2 Typographical properties

3.2.1 Use of capital letters

The text combines capital and miniscule letters. Capital letters are used in different cases:

- The first letter of each word in the titles (e.g. Job Set-up Information, Fault Confirmation, Procedure, etc.) and subtitles (Safety Precautions, Aircraft Maintenance Configuration, etc.). Using a capital letter at the beginning of each word makes easier the information locating. Furthermore, contrary to the miniscule characters, capital letters are more readable from a certain distance (Smith & Mosier, 1986).

- To indicate tools and equipment even though they are cited in the text of the procedure (e.g. « Lock the impeller with the DISASSEMBLY TOOL »)

- To indicate information about hygiene and safety (WARNING, CAUTION, NOTE)

Those capital letters are used to attract the user attention (Smith, Mosier, 1986), particularly when the operator doesn't read entirely the procedure in the maintenance documentation.

3.2.2 Use of bold characters

The use of bold characters to mark information is better than the use of capital letters in information retrieval task (Foster, 1979). For Rivlin et al. (1990) cited by Caro and Bétrancourt (2001), bold character is one of the most salient means to point up an element. This mean is rarely used in the maintenance documentation (only for the column titles in the tables).

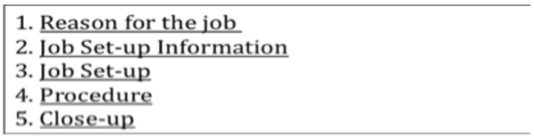
REFERENCE	QTY	DESIGNATION
No specific		Warning notice
MG174-04	1	DISASSEMBLY TOOL

Table 1 - use of bold character

3.2.3 Underlining

In the aircraft maintenance documentation, the underlining is used :

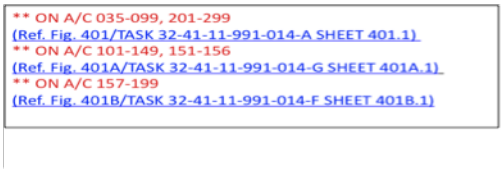
- To accentuate some elements as WARNING, CAUTION, NOTE, or section title of the document. Indeed, the maintenance documentation is divided into different sections corresponding to the different information the operator needs before the execution of the maintenance operation.



1. Reason for the job
2. Job Set-up Information
3. Job Set-up
4. Procedure
5. Close-up

Figure 5 Use of underlining (1)

- To indicate hypertext links: within the documentation, the main procedure can include some subtasks available through hyperlinks. Hyperlinks are also used to view the figure corresponding to a procedure. Those hyperlinks allow saving space on the page. Indeed, a given procedure can concern different aircraft versions or models. In order to not encumber the page with the different figures corresponding to different models, hyperlinks let the user opening a little window to visualise the figure required.



** ON A/C 035-099, 201-299
(Ref. Fig. 401/TASK 32-41-11-991-014-A SHEET 401.1)
** ON A/C 101-149, 151-156
(Ref. Fig. 401A/TASK 32-41-11-991-014-G SHEET 401A.1)
** ON A/C 157-199
(Ref. Fig. 401B/TASK 32-41-11-991-014-F SHEET 401B.1)

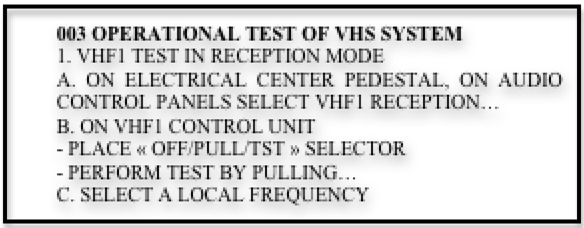
Figure 6 Use of underlining (2)

3.3 Colours

The colours used take into account on the one hand the standard and on the other hand the aeronautical norm ATA (Air Transport Association). Thus, the orange letters are used for the cautions, the red letters for the warnings, the blue for the hyperlink and the yellow underlining to indicate modification in the procedure. In all, five colours are used in the maintenance documentation. It is in accordance with the document ergonomic recommendation, which advises not to use more than five colours in a document to avoid user difficulty in reading. Moreover, a regularity (use of the same colour to indicate the same type of information) can be noted in the use of colours in the maintenance documentation: the red to indicate a danger (warning), the yellow for the attention (caution), the black for normal information (i.e. the procedure) and the blue for the hyperlink

4 Signals rarely or not used in the maintenance documentation

- Indentation: some maintenance documentations don’t use the indentation to organise information. The text is formed by a left centred block without any space between the different sections.



003 OPERATIONAL TEST OF VHS SYSTEM
1. VHF1 TEST IN RECEPTION MODE
A. ON ELECTRICAL CENTER PEDESTAL, ON AUDIO CONTROL PANELS SELECT VHF1 RECEPTION…
B. ON VHF1 CONTROL UNIT
- PLACE « OFF/PULL/TST » SELECTOR
- PERFORM TEST BY PULLING…
C. SELECT A LOCAL FREQUENCY

Figure 7 Lack of indentation in the procedure

- Italic characters: the aircraft maintenance documentation never contains any italic character to distinguish information.

In addition to the ergonomic inspection, a field study allows having more contextual information about the use and the role of signals in the aircraft maintenance documentation. Some results of a field study are presented below.

5 Information from a field study about the role of signals in the aircraft maintenance documentation use

5.1 Presentation of the study

In 2008, we lead an explorative field study with the maintenance operators to understand the use of the aircraft maintenance documentation at the work place. The aim of the study is to grasp the work practice and to analyse the use, the context and the use conditions of the maintenance manual. The study involved 13 operators (11 maintenance operators, 1 foreman and 1 quality manager) aged from 22 to 50, with 12 males and 1 female, at three aircraft maintenance and repair centres in Toulouse. An observation grid has been established in order to gather information concerning the study issue. The object is to know when, how, what for the operator uses the documentation, and if he/she doesn't use it, why? Those questions are synthesized on the figure below.

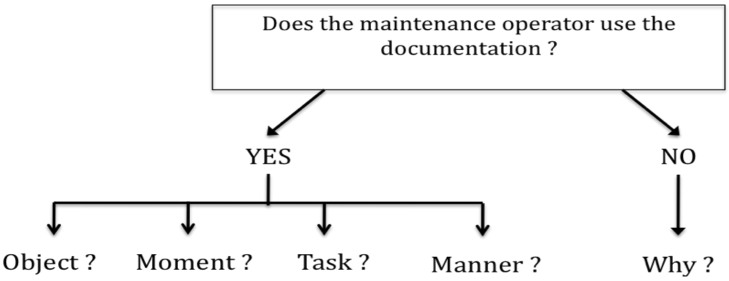


Figure 8 Questions about the use of the documentation by the operator

Interviews with the managers have first been lead and have been completed by the observations of the operators during the task execution. Individual and collective interviews have been carried out with the same operators in order to complete information resulted from the observations.

5.2 Study Results

The results of the study carried out with the operators in the field confirm that the layout and the content of the documentation are the main reasons of the difficulty of the documentation use in the aeronautical maintenance. Indeed, maintenance operators don't use the maintenance documentation. If they use it, they don't read the whole documentation to don't spend too much time to the task execution. Indeed, according to the operators, the maintenance documentation is not easy to use because of different reasons. Those reasons can be divided into two categories depending on the document format (electronic or paper).

5.2.1 Electronic documentation

Reasons of the difficulty can concern the document structure or the document content.

- Document structure

According to the maintenance operators involved in the field study, the maintenance documentation is not very efficient due to the bad information structure and presentation. For instance, elements of the same information are spread over several pages.

- Document content

Other efficient problem is linked to the content: when the user clicks on hyperlink, the system answer doesn't correspond on his expectation. The electronic documentation is efficient for the achievement of some actions (e.g. spare information search from a hyperlink or the reference as key-word) but not to execute maintenance procedure. A map consultation to have a global view of the aircraft is also difficult on a screen documentation consultation. Moreover, information in the electronic documentation can also be insufficient (e.g. the procedure gives the designation of a spare or a tool to be used but any information, such as characteristics and functions about it is in the documentation), or on the contrary in excess when some unnecessary information are presented with the task procedure. According to the operators, it is due to the fact that designers are not in contact with the real work situation, so they could have a bad work situation model. The text they put in the documentation could consequently produce to the user a bad situation model and a bad text understanding. Van Dijk et Kintsch (1983) have proved the role of the situation model on the text understanding. In addition, the procedure is written with complex sentences hard to understand by the operators. In fact, the turn of the phrases in the documentation can sometimes lead up to a bad interpretation of the meaning.

Text in the documentation is monotone because of the lack of text layout, colour, etc. Characters are too small, making reading more difficult and cognitively costly. Figures are difficult to visualise and sometimes don't correspond to the text they are supposed to illustrate

More colours on the figures can help to better understand the procedure because the text is insufficient. Moreover, as the text is not "ventilated", the operators prefer a text organisation in different sections corresponding to each step of the procedure.

5.2.2 Paper Documentation

For the maintenance operators, the paper documentation is the best format to consult a map or to execute technical task, i.e. to follow task procedure. They consider the paper format less efficient to have an overview of the procedure. However, the paper documentation allows them a page-by-page progress in the procedure without a risk of a cognitive disorientation like with the electronic documentation, notably in the case of a long text. Nevertheless, they admit that the ease-of-use of the documentation depends on the content structure and clarity.

6 Conclusion

The field study results and ergonomic inspection show that the maintenance documentation contains some features to distinguish information. However, the means used are not sufficient to make easier the documentation use. The next step would be to take into account the user suggestion and ergonomic recommendation in the documentation design and to evaluate the impact of this modification in the user performance. For maintenance operators, consulting documentation during the task completion constitutes a heavy load. So, according to them, the documentation design has an important role. For operators, an efficient documentation is the one which doesn't require too effort. The availability of clear information and the use of signals could allow having documentation easy to use because of the facilitating of the reading, the information research and consultation.

References

CARO S., BÉTRANCOURT M. (2001). Ergonomie des documents numériques. In *Traité Informatique*, H7 220, Techniques pour l'Ingénieur (TPI) : Paris.

CELLIER J.-M. (2005). Caractéristiques et fonction des textes procéduraux, In D. Alamargot, P. Terrier, & J.M. Cellier (Eds.), *Production, compréhension et usages des écrits techniques au travail* (pp.161-180), Toulouse: Octarès.

CHAPARRO A., GROFF L. S., CHAPARRO B. S., SCARLETT D. (2002). Survey of aviation technical manuals. Phase 2 report : *User evaluations of maintenance documentation*. Federal Aviation Administration.

CHAPARRO A., ROGERS B., HAMBLIN C., CHAPARRO B. (2004). A Comparison of Three Evaluative Techniques for Validating Maintenance Documentation. Final Report.

FOSTER J. J. (1979). The use of visual cues in text. In KOLERS P.A., WROLSTAD M.E., BOUMA H. (Eds.). *Processing of Visible Language,* Vol. 1, p. 189- 203, , London, New York: Plenum Press.

HOBBS A.N. (2000). Maintenance 'error', lessons from the BASI survey. *Flight Safety Australia*, 4, p. 36-37.

LATTANZIO D. PATANKAR K., & KANKI B. (2008). Procedural Error in Maintenance: A Review of Research and Methods. *International Journal of Aviation Psychology*, vol. 18, 1, pages 17-29.

McDonald N., Corrigan S., Cromie S., Daly C. (2000). Safety management systems and safety culture in aircraft maintenance organisations. *Safety Science*, 2000. 34 : 151-176.

Montmollin M. (de). (1997). *Vocabulaire de l'ergonomie.* Toulouse: Octarès Editions, (Coll. Travail).

Rivlin C., Lewis R., Cooper R.D. (1990). Guidelines for screen design. *Blackwell Scientific Publications* LTD.

Sebillotte S. (1991). A task description model based on the operator's objectives. From interview to formalization. *Rapport CENA/INRIA*. Rocquencourt, France : Institut National de Recherche en Informatique et Automatique.

Smith S.L., Mosier J.N. (1986). *Guidelines for designing user interface software*. Bedford : The Mitre Corporation,.

Terrier P., Lemercier C., Mojahid M. (2005). Mise en forme matérielle du texte et traitement de l'information liée à une instruction spécifique : l'effet de mise en acte avec une tâche d'apprentissage, in D. Alamargot, P. Terrier, J.-M. Cellier (Eds.), *Production, compréhension et usages des écrits techniques au travail* (pp.123-143), Toulouse: Octarès.

Van Avermaete J.A.G., Hakkeling-Mesland M.Y. (2001). Maintenance human factors from a European research perspective: results from the ADAMS project and related research initiatives. *15th Annual FAA/TC/CAA Maintenance Human Factors Symposium*, London, England.

Van Dijk T. A., Kintsch W. (1983). *Strategies of discourse comprehension*. San Diego, C.A.: Academic Press.

Natural Language Processing tools applied to Accident Incident Reporting Systems

Eric Hermann Christophe Pimm[1]
(1) CFH, 4 impasse montcabrier 31000 Toulouse
hermann.cfh@orange.fr

Abstract. The human factor field is expected to evolve due to the development of Natural Language Processing tools which allow for new approaches to handle natural language data. Some NLP applications such as machine translation and spell checking are already well known. In the current project, we use NLP methods to facilitate experience feedback in the field of civil aviation safety. The BEA (Bureau of Investigations and Analyses) uses an ECCAIRS repository to integrate data associated with each accident or incident. These data consist of a summary of the circumstances of the accident, written in natural language, and of several fields that are filled with a set of predefined values. NLP methods based on the extraction of information from the summaries can contribute to improve the reliability of the coding, either to facilitate the coding itself or to check the coherence of the data. This approach combines linguistic and statistic treatments. On the linguistic level, a syntactic parser enables us to get relevant terms from the documents; it is the key component of the term extraction method. On the statistical level, correlations are learned from the accident reports to assess the association between the extracted terms and the field values. The system can be easily adapted to other professional fields because it requires light procedures to account for language and concept specificities. NLP tools could be used in order to detect small signals thanks to chronological analyses and detection of perceived risks signals.

Keywords. NLP tools, applied Linguistics, automatic summerization

A framework for corpus-based analysis of the graphic signalling of discourse structure

Martin Thomas, Judy Delin, Rob Waller
Simplification Centre, University of Reading, UK
m.thomas@leeds.ac.uk, judy.delin@roedelin.com, r.waller@reading.ac.uk

Abstract This paper describes a corpus-based approach to the analysis of graphic text signalling in complex information documents. To make the task of populating the corpus tractable, we have developed software to automate as much of the annotation process as possible. OCR output is first obtained in OpenDocument format. This is post-processed semi-automatically to generate stand-off XML annotations following the GeM model (Henschel, 2003). These generated layers describe the content and layout of the document. This information is augmented with functionally-oriented descriptions and RST analyses (Mann and Thompson, 1987). Together these annotations support empirical research into the relationship between the things that are said in documents and the linguistic and graphic resources used to express them. Such research might inform the evaluation of existing documents and the design of new ones.

Keywords: corpus linguistics, discourse organisation, document design

1 Background and motivation

This paper describes a corpus-based approach to the analysis of graphic text signalling in complex information documents from the financial services industry. These documents are products of corporate and regulatory processes, and include content from marketing, customer experience, product technical, administration, legal and compliance specialists. Some are produced by automated systems that dynamically merge customer data with other content. They are influenced by external factors such as government regulation, industry codes, and corporate branding. They sometimes have multiple audiences (for example, customers, intermediaries, and administrators). This complex provenance means that these documents are particularly challenging to analyze, but it also means that a sound analysis will provide valuable insight for the organisations concerned, in terms informing both the evaluation and improvement of existing documents and the design of new ones.

To make the task of populating the corpus tractable, we have developed software to automate as much of the annotation process as possible. We do this by post-processing the output of Optical Character Recognition (OCR). As we have yet to populate the corpus, the focus here will be to document our approach, highlighting our aims and motivating the corpus design. We conclude by giving a taste of the kinds of phenomena which might be investigated on the basis of this design. In parallel with designing the corpus, we have been developing a web-based interface to support querying the data and to present results. Describing this in detail falls beyond the scope of the present paper.

While linguistic accounts of discourse structure differ in important ways, they essentially involve the segmentation of text and the assignment of discourse relations between the resulting segments (see, for example, Grosz and Sidner, 1986; Mann and Thompson, 1987; Martin, 1992). In some instances, surface features may signal these relations. However, as Sporleder and Lascarides (2006, p.371) point out: 'While rhetorical relations are sometimes signalled lexically by discourse markers [...] such as *but*, *since* or *consequently*, these are often ambiguous, either between two or more discourse relations or between discourse and non-discourse usage.' They note that '*since* can indicate either a temporal or an explanation relation' and that *yet* may signal a CONTRAST relation or else be used as a synonym of *so far*, 'with no function with respect to discourse structure whatsoever' (2006, p.371). Moreover, in many cases, relations are not signalled: 'roughly half the sentences in the British National Corpus lack a discourse marker entirely' (2006, p.371). Similarly, in a study of two corpora, one containing task-oriented dialogues, the other a collection of articles from the Wall Street Journal, Taboada (2006) found that, on average, between 60 and 70% of rhetorical relations were not signalled. Interestingly, she also found that some relations are more often signalled than others: in the case of the newspaper corpus, the 'signalling level' for different relations ranged from 4 to 90% (Taboada, 2006, p.587).

While the contribution of graphic resources to the structuring of discourse has long been acknowledged from the perspectives of typographers and information designers (e.g. Hartley and Burnhill, 1977; Waller, 1987), linguists enagaged in discourse analysis (e.g. Bernhardt, 1985; Lemke, 1998) and those working on natural language generation (e.g. Bouayad-Agha, Scott and Power, 2001), there is a gap between this acknowledgement and the kind of corpus-based research that has yielded insights such as those cited above. This is not to say that the need for such work has gone unnoticed. Bateman, Delin and Henschel (2002a, p.8) point out that, in the case of multimodal meaning making, 'we lack convenient cut-off criteria such as "grammaticality" that can be interrogated'. As such, in purely descriptive terms, empirical work is arguably even more critical for building the conceptual scaffolding for multimodal analysis than is the case for linguistic analysis. Moreover, as with language, corpus data can help distinguish between those graphic phenomena that are at all frequent and those which constitute anomalies. While, in some contexts, such anomalies may be of interest in themselves, there is a risk that according them undue significance, as may happen when interpreting hand-picked examples, skews the overall picture. In terms of prescription, while the very lack of a grammar might mean we lack criteria for establishing whether a given instance is well-formed and, as such, run the risk of learning from bad examples, it would nonetheless seem useful to be able to identify the different options selected by document designers in realizing a common set of rhetorical structures.

We take graphic signals to include resources such as layout and spacing, the typographic realization of verbal elements, and non-verbal devices such as bullets, ticks and crosses. In addition to these features of the surface realization of document elements, texts in the corpus are annotated in terms of Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) and a number of other functional labels are applied to segments.

The separation of graphic and linguistic realization from functional descriptions allows investigation of phenomena such as graphic complexity and pacing. We can ask whether variation in graphic realization at a given point is motivated, perhaps by signalling structure or a change in voice. Identifying unmotivated complexity can be seen as a step towards finding opportunities to simplify graphic communications. By supporting the retrieval of instances in which rhetorical structure does, or does not, correspond with artefact stucture (Waller, 1987, p.179), it also enables us to consider the chunking of information in documents.

Apart from establishing, in broad terms, the ways in which graphic signalling is used, a corpus-based approach allows us to ask more specific questions. For example, we might test the hypothesis that graphic signalling is ambiguous, albeit in ways which differ from lexical signalling. Similarly, we might also investigate whether, as seems likely, the distribution of graphic signals differs, in terms of discourse relations, from the distribution of lexical signals. For example, intuitively, bullets might support LIST or JOINT relations, combinations of ticks and crosses might signal ANTITHESIS or CONTRAST relations, and enumeration might be a natural cue for SEQUENCE. Our corpus currently remains too small to make general claims about the role of graphic signalling of discourse relations. However, as we will suggest in section 4, analysis of documents through the framework of the multi-layered annotation described here seems to offer a means of detecting consistency, or its absence, in the design of a given document.

2 Corpus design

In the corpus, each document will be represented at three levels: 1) metadata about the document; 2) transcribed and annotated document content; and 3) a high-resolution facsimile of each page.

Document metadata will facilitate comparison of sets of documents across various dimensions. These will include the sector and brand from which each document has been taken, the genre to which it belongs (e.g. bill, brochure, form, letter), the topic covered (e.g. insurance, pensions, tax) and the date it was produced. Where relevant, we will identify the relationship between each document and others included in the corpus (in terms of document versions and in relation to steps in business processes). Where available, we will also record data relating to business process errors associated with the document and any existing expert evaluations.

The annotation of specific document elements will be discussed in greater detail in the next section. In brief, it allows us to compare what is said, in rhetorical and functional terms, with how it is said, in terms of graphic and linguistic realizaton. Genre, for which a characterization is given to a whole document, can also be associated with document parts: thus instances of embedded genres, such as a form within a brochure, may be described and retrieved.

3 Corpus annotation

The corpus annotation is based on the GeM scheme described comprehensively by Henschel (2003). The scheme implements stand-off annotation in XML layers. These include *base*, *layout* and *rhetorical structure* (henceforth *RST*). We have converted the original GeM DTDs into RelaxNG[1] schemas.

This stand-off approach provides an elegant way in which to accommodate the annotation of cross-cutting and overlapping phenomena. It also supports the straightforward addition of new annotation layers, which allows us to record additional descriptions of document segments, such as speaker, intended audience, or communicative intention.

3.1 Capturing data

Leech (1991, p.10) notes that, during the 1980s, OCR freed 'corpus compilation from the logjam of manual input'. Approaches which seek to represent the graphic as well as linguistic realization of texts inevitably face an even more formidable logjam if manual input is the only means of capturing data. In addition to making the transcription and annotation task tractable, exploiting recent developments in OCR technology to capture information about graphic realization has the additional benefit of excluding judgements about rhetorical relatedness from consideration when describing the visual segmentation of document elements. Such influence, to which human annotators may unavoidably be susceptible, risks introducing circularity into subsequent analysis of information chunking.

Given that we wish to retain as much information as possible about the graphic realization of each document, including the layout and placement of elements as well as specifics such as the size, weight and colour of individual characters, the format in which we choose to save the OCR output should be adequately descriptive. The processes and tools we have implemented to support corpus development will be available for the ongoing expansion of the corpus. The format chosen should therefore be sustainable: it should not be liable to withrawal or obsolescence. As such, it should be a documented open standard. Finally, the format must afford automated processing downstream. Given that our target corpus annotation scheme is XML-based, using an XML format to store the raw OCR output offers a particularly good fit. The format which currently best meets these three criteria is OpenDocument[2], which is published as an ISO standard and can be used with a wide variety of office suites, including OpenOffice and Microsoft Office.

Support for OpenDocument in OCR software is increasing. Where it is not available for saving OCR output, this may first be saved as a Microsoft Word document and then converted without loss into OpenDocument using either Microsoft Word or OpenOffice.

[1] http://www.oasis-open.org/committees/relax-ng/spec-20011203.html

[2] http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office

3.2 Post-processing OCR output with XSLT

Using XSLT we can transform OCR output in OpenDocument format automatically into something similar to the *base* and *layout* layers defined in the GeM scheme. Essentially, this involves parsing the source document to assign each element a unique identifier and then reverse engineering the style information carried by each element to conform to the GeM formalism. The other automated intervention is to segment running text at sentence boundaries. This supports RST annotation and the future addition of new layers accommodating other linguistic analyses.

At this point, some rather fundamental differences between the underlying motivations behind OpenDocument and the GeM scheme become evident. OpenDocument is intended to specify document rendering from an authoring perspective, GeM describes the presentation of documents as end products. While it records the layout of elements in relation to a grid-like structure, OpenDocument is not required to record the placement of elements in relation to vertical space. Thus we do not know the extent of each row in the grid. Neither is there an explicit indication of the page on which a given element is placed. Finally, while column widths are retrievable, the placement of individual elements on a given line is not specified: it seems the text is poured into a container and the editing software decides its precise placement. For our purposes, the most significant consequence of this is that, while we can automatically retrieve information about the column and row in which an element is placed, if we wish to indicate page boundaries these must be added manually.

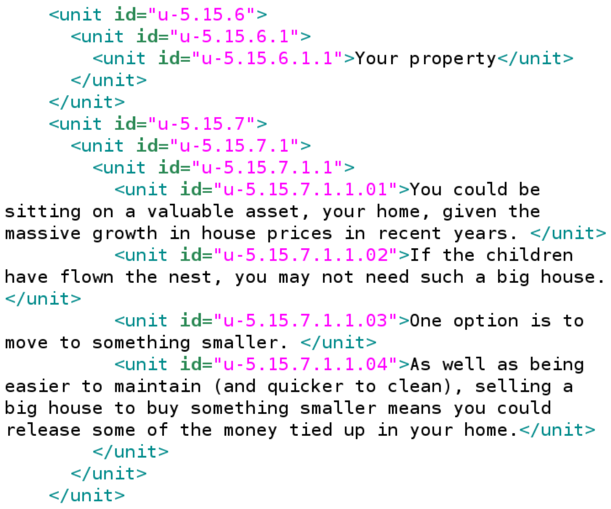
In order to establish the robustness of our approach, we have tested our transformations on a suite of documents containing various graphic features, including complex layouts, tables and lists, for which various routes were taken to arrive at their OpenDocument form.[3] Success varies in terms of representation of layout segmentation, document flow and the capture of features such as lists. However, this variation is a function of the OCR process and the representation of OCR output in OpenDocument, rather than of our post-processing. One advantage of using a pivot format, such as OpenDocument, is that manual tweaks to layout recognition during OCR or improvements in OCR technology can feed directly into our processing model without the need for modification to our tools. Our tools coped well with the range of scenarios with which we have presented them, consistently allocating unique identifiers to elements and retrieving available information about their graphic realization.

3.3 Base and layout annotation layers

The *base* layer carries a transcription of the verbal content of the document. In the case of non-verbal graphics, a verbal description of the element is provided. The other purpose of the base layer is to segment the document into *base units*. Each unit is assigned a unique identifier (see Figure 1). Using this identifier, the base units are cross-referenced by the other layers.

Within the *layout* layer, there are four main sections: 1) layout *segmentation* in which each *layout unit* cross-references one or more base units; 2) *realization* information; 3) an abstract

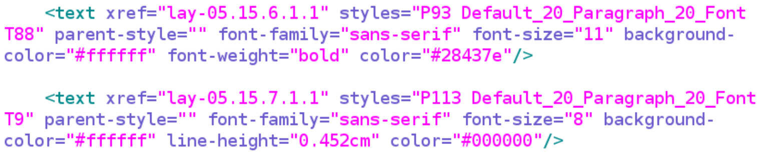
[3] OCR output in various Microsoft Word formats was converted by Microsoft Word and OpenOffice into OpenDocument format. In one case, we reconstituted a document in a word processor by merging OCR output from scanned images made on different scanners.



```
      <unit id="u-5.15.6">
        <unit id="u-5.15.6.1">
          <unit id="u-5.15.6.1.1">Your property</unit>
        </unit>
      </unit>
      <unit id="u-5.15.7">
        <unit id="u-5.15.7.1">
          <unit id="u-5.15.7.1.1">
            <unit id="u-5.15.7.1.1.01">You could be
sitting on a valuable asset, your home, given the
massive growth in house prices in recent years. </unit>
            <unit id="u-5.15.7.1.1.02">If the children
have flown the nest, you may not need such a big house.
</unit>
            <unit id="u-5.15.7.1.1.03">One option is to
move to something smaller. </unit>
            <unit id="u-5.15.7.1.1.04">As well as being
easier to maintain (and quicker to clean), selling a
big house to buy something smaller means you could
release some of the money tied up in your home.</unit>
          </unit>
        </unit>
      </unit>
```

Figure 1: Fragment of base layer annotation

representation of the overall layout of the display, known as the *area model*; and 4) a more concrete description of the *layout structure* of the document, in which layout units are located within the *area model*. Taken together, these four components allow us to build a fairly comprehensive picture of the graphic realization of document elements.



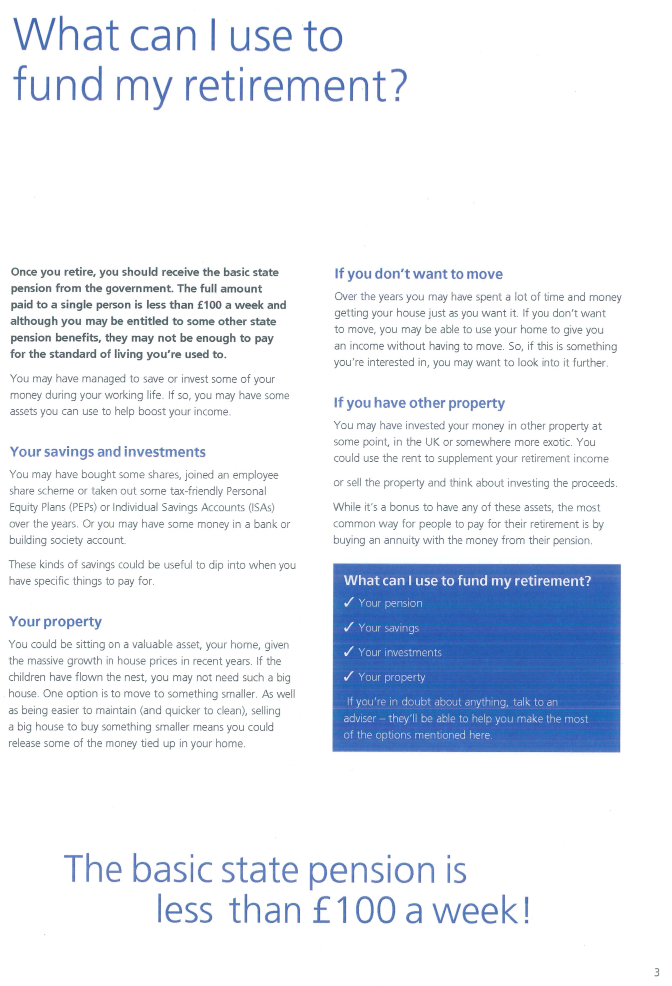
```
    <text xref="lay-05.15.6.1.1" styles="P93 Default_20_Paragraph_20_Font
T88" parent-style="" font-family="sans-serif" font-size="11" background-
color="#ffffff" font-weight="bold" color="#28437e"/>

    <text xref="lay-05.15.7.1.1" styles="P113 Default_20_Paragraph_20_Font
T9" parent-style="" font-family="sans-serif" font-size="8" background-
color="#ffffff" line-height="0.452cm" color="#000000"/>
```

Figure 2: Fragment of layout layer realization annotation

The fragment shown in Figure 2 describes details of the typographic realization of two verbal document elements. The first is realized in 11 point bold type. The hexadecimal RGB value for *color* describes a shade of blue. The second element is smaller, at 8 points, and black. In both cases, the background colour is white. These two descriptions relate to the heading ‘Your property’ and subsequent paragraph at the bottom of the left-hand column in Figure 3. The realization information retrieved from the data in OpenDocument format differs slightly from that specified by GeM. Some additional information is present (e.g. *line-height*), while other parameters are missing (e.g. *justification* and *case*). Our RelaxNG schema for the layout layer defines additional new attributes and relaxes existing requirements as necessary. We do not believe that this involves a significant deviation from the original annotation model.

The fragment of the *area model* shown in Figure 4 describes the overall layout of the page on which our previous example is realized: in a slight deviation from the GeM specification, we allocate an *area-root* element to each page in the document. The model assumes a grid layout in which each *sub-area* represents a row. So, our example describes a display comprising three rows. The first of these consists of a single column. The second has two columns of approximately equal width. The third also has two columns, but in this the case the first column, on the left, is much wider than the second, which just contains the page number, which is barely



What can I use to fund my retirement?

Once you retire, you should receive the basic state pension from the government. The full amount paid to a single person is less than £100 a week and although you may be entitled to some other state pension benefits, they may not be enough to pay for the standard of living you're used to.

You may have managed to save or invest some of your money during your working life. If so, you may have some assets you can use to help boost your income.

Your savings and investments

You may have bought some shares, joined an employee share scheme or taken out some tax-friendly Personal Equity Plans (PEPs) or Individual Savings Accounts (ISAs) over the years. Or you may have some money in a bank or building society account.

These kinds of savings could be useful to dip into when you have specific things to pay for.

Your property

You could be sitting on a valuable asset, your home, given the massive growth in house prices in recent years. If the children have flown the nest, you may not need such a big house. One option is to move to something smaller. As well as being easier to maintain (and quicker to clean), selling a big house to buy something smaller means you could release some of the money tied up in your home.

If you don't want to move

Over the years you may have spent a lot of time and money getting your house just as you want it. If you don't want to move, you may be able to use your home to give you an income without having to move. So, if this is something you're interested in, you may want to look into it further.

If you have other property

You may have invested your money in other property at some point, in the UK or somewhere more exotic. You could use the rent to supplement your retirement income

or sell the property and think about investing the proceeds.

While it's a bonus to have any of these assets, the most common way for people to pay for their retirement is by buying an annuity with the money from their pension.

What can I use to fund my retirement?

✓ Your pension

✓ Your savings

✓ Your investments

✓ Your property

If you're in doubt about anything, talk to an adviser – they'll be able to help you make the most of the options mentioned here.

The basic state pension is less than £100 a week!

3

Figure 3: Facsimile of the display described in Figure 4.

visible on the far right.

```
<area-root width="20.999cm" height="29.699cm">
  <sub-area id="14" style-name="P54" cols="1" hspacing="100"/>
  <sub-area id="15" style-name="Sect4" cols="2" hspacing="4775 4819"/>
  <sub-area id="17" style-name="Sect6" cols="2" hspacing="8272 1735"/>
</area-root>
```

Figure 4: Fragment of layout layer area model annotation

This rather abstract description is populated with layout units in the *layout-structure*. Putting the fragment shown in Figure 5 together with the *area model* reproduced in Figure 4, we see that our heading 'Your property' (*lay-5.15.6*) and the subsequent paragraph (*lay-5.15.7*) are placed last in the first column of the second row in the display (sub-area *15*).

In terms of *layout-structure*, document elements may be related recursively: a column in a row in the grid representing the overall page may contain a table which in turn contains text arranged in columns and rows. In addition to recording tabular layouts, elements displayed as a list in the document are also identifiable as such from the *layout-structure*.

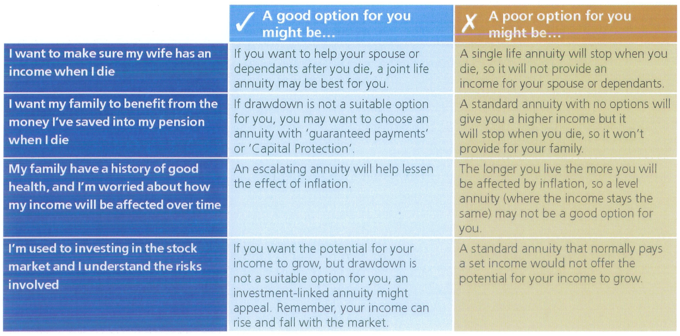
```
<layout-chunk area-ref="15">
  <layout-leaf xref="lay-5.15.1" location="col-1"/>
  <layout-leaf xref="lay-5.15.2" location="col-1"/>
  <layout-leaf xref="lay-5.15.3" location="col-1"/>
  <layout-leaf xref="lay-5.15.4" location="col-1"/>
  <layout-leaf xref="lay-5.15.5" location="col-1"/>
  <layout-leaf xref="lay-5.15.6" location="col-1"/>
  <layout-leaf xref="lay-5.15.7" location="col-1"/>
  <layout-leaf xref="lay-5.15.8" location="col-2"/>
  <layout-leaf xref="lay-5.15.9" location="col-2"/>
  <!-- ... -->
  <layout-leaf xref="lay-5.15.19" location="col-2"/>
</layout-chunk>
```

Figure 5: Fragment of layout structure annotation

3.4 RST and other semantically-oriented layers

Adding functional labels to segments in the *RST* and other semantically-oriented layers must be done manually. This said, by automating the generation of skeleton files, the human annotator is largely relieved of the task of managing cross references across annotation layers.

Despite adopting RST, the GeM project identified several problems with implementing it in multimodal analysis (see, for example, Bateman, Delin and Henschel, 2002b; Bateman, 2008). Perhaps the most fundamental modification to RST as implemented in GeM is the generalization of the RST sequentiality assumption to allow relations to document parts which are adjacent in any direction (Henschel, 2003, p.15).



	✓ A good option for you might be...	✗ A poor option for you might be...
I want to make sure my wife has an income when I die	If you want to help your spouse or dependants after you die, a joint life annuity may be best for you.	A single life annuity will stop when you die, so it will not provide an income for your spouse or dependants.
I want my family to benefit from the money I've saved into my pension when I die	If drawdown is not a suitable option for you, you may want to choose an annuity with 'guaranteed payments' or 'Capital Protection'.	A standard annuity with no options will give you a higher income but it will stop when you die, so it won't provide for your family.
My family have a history of good health, and I'm worried about how my income will be affected over time	An escalating annuity will help lessen the effect of inflation.	The longer you live the more you will be affected by inflation, so a level annuity (where the income stays the same) may not be a good option for you.
I'm used to investing in the stock market and I understand the risks involved	If you want the potential for your income to grow, but drawdown is not a suitable option for you, an investment-linked annuity might appeal. Remember, your income can rise and fall with the market.	A standard annuity that normally pays a set income would not offer the potential for your income to grow.

Figure 6: Table describing good and poor options

One particular challenge presented by the material we have annotated is the presence of tables. Clearly, these support rhetorical relations between elements along more than one spatial dimension. To accommodate this, we allow segments to participate in more than one RST structure. Thus, considering the example in Figure 6, moving from left to right each of the rows of content presents a conditioning situation (CONDITION) followed by a pair of outcomes, which themselves form a CONTRAST relation. In terms of columns, the cells can be seen as constituting a set of situations in a JOINT relation. The column headings relate to the cells below as PREPARATION. This implementation, in which we allow segments to participate in multiple orthogonal relations, retains the concept of nuclearity central to RST.

Other layers are used to describe the *speaker*, *audience*, communicative *intention* and *genre* of document segments. These are formally less complex than the RST layer. They consist of a series of segments each of which has an *xref* attribute which cross references contiguous *base* units who share a given property or, in the case of audience, properties. Thus, the annotation

```
  <speaker-unit id="s-13.56.2" xref="u-13.56.2 u-14.57.1 u-14.57.2 u-14.58.1 u-14.60.1
u-14.60.2 u-14.60.3 u-14.60.4 u-14.60.5 u-14.60.6 u-14.60.7 u-14.60.8 u-14.60.9
u-14.60.10 u-14.60.11 u-14.60.12 u-14.60.13 u-14.60.14" type="owner"/>
  <speaker-unit id="s-14.60.15" xref="u-14.60.15" type="endorser"/>
  <speaker-unit id="s-14.62.1" xref="u-14.62.1 u-14.62.2 u-15.63.1 u-15.63.2 u-15.64.1
u-15.65.1 u-15.66.1 u-15.66.2 u-15.67.1 u-15.68.1 u-15.68.2 u-15.68.3 u-15.68.4
u-15.68.5 u-15.68.6 u-15.68.7 u-15.68.8 u-15.68.9 u-15.68.10 u-15.69.1 u-15.70.1
u-15.71.1" type="owner"/>
```

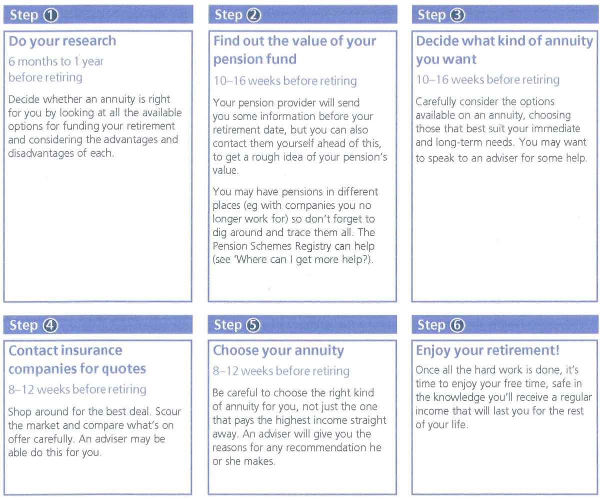
Figure 7: Fragment of *speaker* layer annotation

fragment in Figure 7 shows that a segment spoken by a third-party *endorser* (*s-14.60.15*) is inserted within a context in which the document *owner* is speaking. In this instance, if we compared the *speaker* with the *layout* layer, we would find that this voice is differentiated typographically in terms of placement, size and colour.

4 Conclusions and further work

The corpus is currently too small to support general claims about graphic signalling of discourse structure and, having put the processes and tools in place, the most pressing future work involves its population. This said, the pilot annotations performed so far have drawn attention to certain features and inconsistencies within individual documents.

All of the examples we have presented here have been taken from a brochure produced by a life insurance company for distribution to people approaching retirement. In many ways it is a successful design. The brochure is aligned with the reader, who it supports through a decision-making process. It terms of information structure, it is well-paced. In most cases, each display presents a topic coherently and topics do not run across pages. In terms of our annotation model, this informal judgement may be supported by comparing the RST and layout layers. The combination of good spacing and short line length maintains legibility, despite much of the type being relatively small.



Step ①

Do your research

6 months to 1 year
before retiring

Decide whether an annuity is right for you by looking at all the available options for funding your retirement and considering the advantages and disadvantages of each.

Step ②

Find out the value of your pension fund

10–16 weeks before retiring

Your pension provider will send you some information before your retirement date, but you can also contact them yourself ahead of this, to get a rough idea of your pension's value.

You may have pensions in different places (eg with companies you no longer work for) so don't forget to dig around and trace them all. The Pension Schemes Registry can help (see 'Where can I get more help?).

Step ③

Decide what kind of annuity you want

10–16 weeks before retiring

Carefully consider the options available on an annuity, choosing those that best suit your immediate and long-term needs. You may want to speak to an adviser for some help.

Step ④

Contact insurance companies for quotes

8–12 weeks before retiring

Shop around for the best deal. Scour the market and compare what's on offer carefully. An adviser may be able do this for you.

Step ⑤

Choose your annuity

8–12 weeks before retiring

Be careful to choose the right kind of annuity for you, not just the one that pays the highest income straight away. An adviser will give you the reasons for any recommendation he or she makes.

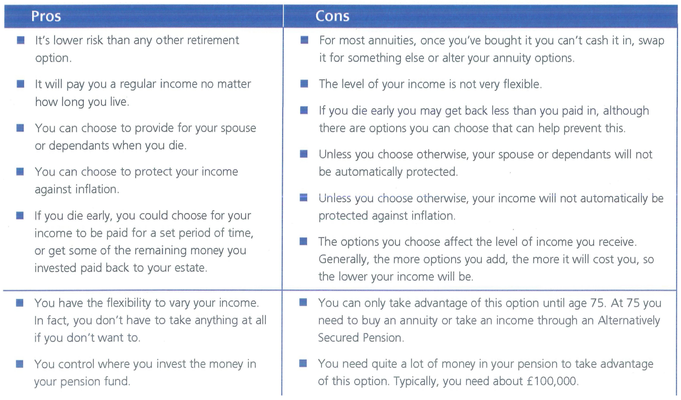
Step ⑥

Enjoy your retirement!

Once all the hard work is done, it's time to enjoy your free time, safe in the knowledge you'll receive a regular income that will last you for the rest of your life.

Figure 8: Your next steps

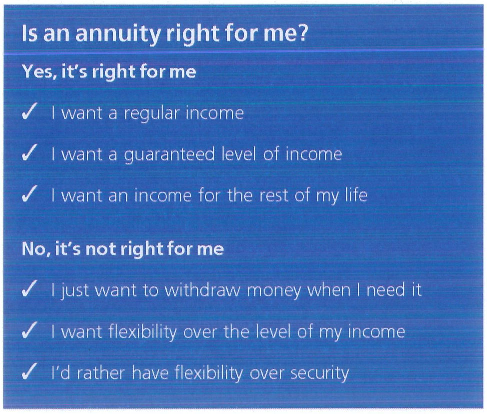
However, with regard to the graphic signalling of specific rhetorical relations, a mixed picture emerges. On the one hand, we note that the one clear instance of a SEQUENCE relation is marked by enumeration and that each step is graphically framed (see Figure 8). One the other hand, a combination of ticks and crosses, and bullets of different shapes and colours are used to cue list items at various points in the document. The use of these resources seems to lack consistency.



Pros	Cons
■ It's lower risk than any other retirement option. ■ It will pay you a regular income no matter how long you live. ■ You can choose to provide for your spouse or dependants when you die. ■ You can choose to protect your income against inflation. ■ If you die early, you could choose for your income to be paid for a set period of time, or get some of the remaining money you invested paid back to your estate.	■ For most annuities, once you've bought it you can't cash it in, swap it for something else or alter your annuity options. ■ The level of your income is not very flexible. ■ If you die early you may get back less than you paid in, although there are options you can choose that can help prevent this. ■ Unless you choose otherwise, your spouse or dependants will not be automatically protected. ■ Unless you choose otherwise, your income will not automatically be protected against inflation. ■ The options you choose affect the level of income you receive. Generally, the more options you add, the more it will cost you, so the lower your income will be.
■ You have the flexibility to vary your income. In fact, you don't have to take anything at all if you don't want to. ■ You control where you invest the money in your pension fund.	■ You can only take advantage of this option until age 75. At 75 you need to buy an annuity or take an income through an Alternatively Secured Pension. ■ You need quite a lot of money in your pension to take advantage of this option. Typically, you need about £100,000.

Figure 9: Pros and cons

It might be argued that the ticks in Figure 3 carry a positive value, which justifies their use in place of the circular bullets used elsewhere in the brochure. However, the lack of any graphic differentiation between the lists of pros and cons in Figure 9 might be seen as a missed opportunity. Finally, ticks cue both the positive and the negative values in the two lists shown in Figure 10. Here the rhetorical relation CONTRAST would appear not to be supported by the graphic signalling: indeed, one might argue that it is undermined by it. It is unclear why a graphic approach consistent with that already presented in the table reproduced in Figure 6 was not deployed in either of these latter two cases.



Is an annuity right for me?

Yes, it's right for me

✓ I want a regular income

✓ I want a guaranteed level of income

✓ I want an income for the rest of my life

No, it's not right for me

✓ I just want to withdraw money when I need it

✓ I want flexibility over the level of my income

✓ I'd rather have flexibility over security

Figure 10: Is an annuity right for me?

In sum, then, comparing the the layout and RST annotation layers reveals instances in which a common graphic realization serves different rhetorical functions and in which different realizations support a common rhetorical function. If this kind of analysis is revealing when applied to a single document, it seems likely that the application of this approach to a broader set of questions and much larger collections of comparable documents will yield significant new insights.

References

Aijmer, K. and Altenberg, B. (eds) (1991). *English Corpus Linguistics: Studies in Honour of Jan Svartik*, Longman, Harlow.

Bateman, J. A. (2008). *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*, Palgrave Macmillan, Houndmills.

Bateman, J., Delin, J. and Henschel, R. (2002a). Multimodality and empiricism: Methodological issues in the study of multimodal meaning-making. GeM project report.
URL: *http://www.purl.org/net/gem*

Bateman, J., Delin, J. and Henschel, R. (2002b). XML and multimodal corpus design: experiences with multi-layered stand-off annotations in the GeM corpus, *Proceedings of the LREC'02 Workshop 'Towards a roadmap for multimodal language resources and evaluation'*.

Benson, J. and Greaves, W. (eds) (1985). *Systemic perspectives on discourse*, Vol. 2, Ablex, Norwood, NJ.

Bernhardt, S. A. (1985). Text structure and graphic design: the visible design, *in* Benson and Greaves (1985), pp. 18–38.

Bouayad-Agha, N., Scott, D. and Power, R. (2001). The influence of layout on the interpretation of referring expressions, *in* Degand, Bestgen, Spooren and van Waes (2001), pp. 133–141.

Degand, L., Bestgen, Y., Spooren, W. and van Waes, L. (eds) (2001). *Multidisciplinary Approaches to Discourse*, Stiching Neerlandistiek VU, Amsterdam.

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions, and the structure of discourse, *Computational Linguistics* **12**(3): 175–204.

Hartley, J. and Burnhill, P. (1977). Understanding instructional text: Typography, layout, and design, *in* M. Howe (ed.), *Adult Learning*, Wiley, London, pp. 223–247.

Henschel, R. (2003). *GeM Annotation Manual Version 2*, GeM Project.
URL: *http://www.purl.org/net/gem*

Leech, G. (1991). The state of the art in corpus linguistics, *in* Aijmer and Altenberg (1991), pp. 8–29.

Lemke, J. (1998). Multiplying meaning: Visual and verbal semiotics in scientific text, *in* Martin and Veel (1998), pp. 87–113.

Mann, W. and Thompson, S. A. (1987). Rhetorical structure theory: A theory of text organization, *Technical report*, Information Sciences Institute, Los Angeles.

Martin, J. (1992). *English Text: System and structure*, John Benjamins, Philadelphia.

Martin, J. and Veel, R. (eds) (1998). *Reading Science: Critical and functional perspectives on discourses of science*, Routledge, London.

Sporleder, C. and Lascarides, A. (2006). Using automatically labelled examples to classfiy rhetorical relations: an assessment, *Natural Language Engineering* **14**(3): 369–416.

Taboada, M. (2006). Discourse markers as signals (or not) of rhetorical relations, *Journal of Pragmatics* **38**: 567–592.

Waller, R. (1987). *The Typographic Contribution to Language*, PhD thesis, University of Reading.

Analysis of Textual Data, some topological methods for studying text structure indicators: the case of Latin historic narratives[1]

Dominique Longrée (1), Sylvie Mellet (2)

(1) LASLA, Université de Liège
Dominique.Longree@ulg.ac.be
(2) BCL, Université Nice Sophia Antipolis, CNRS
Sylvie.Mellet@unice.fr

Abstract: Our research uses the analysis of a Latin historical corpus to study the indicators structuring literary texts in general, and focuses on methods which are valid for any text of some length. Our basic assumption is the following: such texts include complex multilevel structures (i.e. those calling upon lexis, semantics, morphology, syntax...) which function as heterogeneity indicators (progression to a new episode, focalisation on a new point of view, insertion of a reported speech, etc.). Additionally, the recurrence of these structures is a factor in textual cohesion and helps the reader's memorization. Under certain conditions, they function as topological "motives" marking the linear progression of the text and ensuring textual unity. We are developing new tools for detecting such "motives" and for allowing structural comparison in the perspective of a contrastive corpus study (contrasts between genres, authorial styles, etc.). Our methods are based on mathematical models (neighbourhoods, bursts) and combine a small-scale qualitative approach with large-scale quantitative analysis.

Keywords: text structure, "motives", topological models, quantitative methods, Latin historical corpus.

The framework of the present study is a long term research work on the textual elements enabling us to characterise a text or group of texts with a view to automatic classification, and, at the same time, to signal the way they are organised with a view to an equally automatic structuring. This presentation aims at demonstrating the interest and the effectiveness of an approach which intersects classical philology and textual data analysis. By leaning heavily on a corpus of Latin historians whose works have been lemmatised and morphosyntactically labelled,[2] we present here some of the methods we have developed which have in common that they combine an enriching small-scale qualitative analysis with

[1] This research is supported by Wallonie-Bruxelles International, by the Belgian Funds for Scientific Research, by the French Ministry of European and Foreign Affairs, and by the French Ministry of Higher Education and Research as part of the Hubert Curien cooperation programs.

large-scale quantitative methods in order to let the text structures show up by themselves, in an objective and endogenous way. We analyse the results already obtained as well as the difficulties encountered. We hope to convince in terms of the necessity of tightly linking up a qualitative approach and a quantitative data processing in order to detect the signals of textual organisation and to reach a fuller understanding of their role. This presentation on methodological practices, which is mainly carried out from the point of view of the analyst, should nevertheless provide pertinent food for thought for anyone who is interested in the cognitive process of textual reading.

1 The problematic

Classical philology has for a long time recognised, more or less intuitively, numerous text organisation signals and has made an inventory of them, on the one hand – and mainly – with an eye to characterising a literary genre or an author's style, and on the other hand for teaching purposes (the indicators in question providing learners with reading instructions). Amongst the main indicators catalogued in Latin narrative prose, we will highlight:
- sequences of verb tenses, with breaks or in series, such as the opening imperfect at the beginning of a chapter, the imperfects with a break with a series of narrative tenses, a series of perfects[3] or historical presents;
- lexical elements, such as certain adverbs (for example *repente*, *subito* 'suddenly', 'abruptly');
- syntactical structures, most often combined with preceding elements: the most common amongst them are the so-called 'linking clichés', based on a combination of syntactic structures such as participial or temporal clauses with a limited stock of lexical items (*Quibus rebus cognitis* 'Those things being known', *Dum haec in Gallia geruntur* 'whilst these events are taking place in Gaul', *Quod ubi animaduertit* 'When he had noticed that'); other indicators are, for instance, accumulations of temporal clauses at the head of a sentence or series of historical infinitives in combination with sequences of paratactic main clauses.

Nevertheless these observations quickly run up against their limits:

- The indicators pointed out are seen as recurrent signs, but are not really analysed in terms of their function as organisers of the text's structure;
- Their respective weight is not assessed;
- Traditional analysis offers only a few tools to make exactly clear the interactions between the indicators;
- Until very recently, estimating their contribution to variations in genre and style remained intuitive and approximate.

However, over the past few years, the « Tekstcohesie » team of the Dutch research group OIKOS has provided new insights into this question by using the concepts of discourse modes and bases (i.e. narration time or reference time). Within this framework, the discourse modes are defined *a priori*[4]. Latin texts are segmented according to the chosen set of modes.

[2] The corpus offered by the University of Liège's 'Laboratoire d'Analyse Statistiques des Langues Anciennes' (LASLA) textual data base, and here we work in particular on Caesar 's *Civil Wars* Book 2 and on Book 12 of Tacitus' *Annals*.

[3] The Latin perfect covers the use of the French past historic and present perfect.

[4] See Adema, 2009: for instance, "The *description mode*, consists of what every reader would indeed call description: sequences in which the narrator gives the visible characteristics of an object or location" or "The registering mode consists of those states of affairs which describe what the narrator experiences in his

Regarding as one unit the discontinuous segments belonging to the same discourse mode and base, the researchers try to associate with each unit some typical linguistic features by studying, for instance, tense uses (Kroon, 2007; Adema, 2007, 2008, 2009) or referential choices (Kroon, 2009). The features typical of each discourse mode appear in a way to be text structure indicators, and by taking into account features belonging *a priori* to discourse modes other than the one considered, one can evaluate how homogeneous the particular discourse mode is. However, the main purpose of this research is either to describe linguistic phenomena (such as the anaphora) or to pinpoint, on a text linguistic level, differences between texts of different genres and/or from different authors.

Our approach to the question is quite different. On the one hand, we attempt to limit the impact of the linguist's subjectivity: even if we are compelled to select the linguistic phenomena we regard as relevant for the organisation of the text, we want to avoid defining *a priori* sets of linguistic features associated with a given kind of discourse and to project this *a priori* assumption onto the facts. We will develop endogenous exploratory methods and let the texts “speak for themselves”. On the other hand, we regard the selected features as text structure indicators that can be spotted as the text goes by; for this reason, when we describe their functioning, our main concern is precisely to take into account this text linearity towards which the very function of these indicators is to bring together discourse heterogeneity and text continuity.

Therefore we will try to go past the limits mentioned above and to present tools which first of all allow for a specification of the functional convergences between several indicators of textual structuring, and then a linking of the latter to the internal structure of each text and to its narrative progression whilst taking completely into account the writing specificities of each author.

There are three of these conceptual and methodological tools: we will call on the notions of motif, neighbourhood and bursts.

2 The motif

The notion of motif has been conceived of as a means of conceptualising the multi-dimensionality (or multi-level character) of some textual indicators which call upon lexis, grammatical categories and syntax at the same time. In a formal way the motif is defined as an ordered subset of the textual ensemble (T), formed by the recurring combination of n elements of (T) provided with its linear structure. Thus, if the text is formed by a certain number of occurrences of elements A, B, C, D and E, a motif can be the recurring micro-structure ACD or AAA, etc., without here prejudging the nature of the elements A, B, C, D and E in question: the motif is only the framework – or the envelope – accommodating a range of parameters to be defined and capable of characterising the diverse texts of a corpus, or even the different parts of a text.

If the defining linguistic units of the motif are of an exclusively lexical nature, the latter comes appreciably close to what André Salem recognized through the name of ‘repeated segments’ or to the units which, in another field of application, are called ‘phraseological units.’ But the advantage of the notion of the motif is that it embraces on the one hand the heterogeneity of the components (a motif can thus be formed jointly with lexical and

immediate environment, i.e. the time and place of narration”.

grammatical parameters, or with phonological or metrical structures), and on the other the presence of variables within itself.

We can illustrate these properties through the linking cliché *Dum haec in Gallia geruntur*[5] which is characterised by a syntactical structure (a temporal clause introduced by the conjunction *dum*, always followed by the present indicative), by an actantial schema (neutral plural anaphoric as the subject of the verb *gero* in the passive voice) and by its frontal position in the sentence. The locative is most commonly required as well (it is indispensable at the announcement of a change of scene location); if it is absent the motif simply introduces a second aspect of the same episode. The lexical stock is extremely limited, as can be seen in the paradigm below which takes in all the occurrences in Caesar and Tacitus, but it nevertheless allows for variation.

We thus here typically have what we call a ‘framing motif’ (fr.: “motif **cadratif**”), borrowing terminology from Michel Charolles which we have adapted a little to our corpus and to the Latin tradition to designate the ensemble of sentential structures in which the writer straightaway sets down the circumstantial settings (fr.: “**cadre** circonstanciel”) of the narrated event.[6]

We can thus imagine the benefits of being able to detect automatically, in a digitalised corpus, this paradigm of utterances. The first step of course consists of having available a corpus which is annotated at several levels of analysis; the XML format has made this technically possible; the results given by lemmatisers and syntactical analysers provide lexical and grammatical labels which are pretty much satisfactory; on the other hand, the tools for the other levels of annotation are still either very disappointing (at the semantic level) or in their infancy (phonological, prosodic and pragmatic levels). And, above all, we are cruelly lacking in high-performance tools to read and use the multi-level annotation and to respond to complex requests which would be capable of making our motifs come to the fore in an automatic manner.[7]

Therefore we are only able to take into account more simple indicators; but we will nevertheless make every effort to evaluate the convergence of their functioning within the textual framework; we will provisionally leave to one side the lexicon and focus our study on the possible interactions between temporal sequences and purely syntactical, framing or appendage motifs; more specifically we will use four parameters: series of perfects, series of presents, combinations of participial clauses (ablative absolutes) and the *cum* + subjunctive clauses (‘while’) at the head of a sentence and the same combinations at the end of a sentence.

Our philological knowledge of the Latin historians confirms that these parameters have key roles in the narrative dynamic of the texts, but we do not project any *a priori* assumptions onto the possible correlations between them, onto their distributions along the text stream, or

[5] Literally ‘Whilst these things are being done in Gaul’, i.e. ‘Whilst these events tare taking place in Gaul’; e.g. in Caesar: *Dum haec ad Ilerdam geruntur [...], Dum haec in Hispania geruntur [...], Haec dum apud hostes geruntur [...], Dum ea [Ø] geruntur [...], Dum haec per provincias a Vespasiano ducibuque partium geruntur [...].* A variant is found in Tacitus, *Hist*, 2, 87 in which the verb *geruntur* is replaced by the verb *parantur* (‘are prepared’ instead of ‘are being done’).

[6] This type of structure is, in the philological tradition, the opposite of what are termed ‘sentences with an appendage’ in which the historian adds, after the main clause, additional circumstantial elements extending a sentence which seemed to have terminated.

[7] One pathway, amongst others, has been opened up Loiseau’s CorpusReader project (Loiseau, 2007), but the performances still remain disappointing.

onto their specific structuring functions and values. In order to study these various phenomena, we will evaluate two large-scale quantitative methods – the neighbourhood method and the bursts method – which, at both the micro- and macro-structural levels, react differently to the frequency and distribution of the chosen parameters. These distributions will be analysed in the two texts we selected (*Civ2* and *Ann12* – see footnote 1): the parameters are found in both texts, but distributed and used differently, which will enable us, through a comparison of the two texts, to prove the heuristic value of the methods used.

3 The neighbourhood

3.1 Definition and application

The text can be conceived of as a linear structure composed of an ensemble of linguistic events.[8] Considering each of these linguistic events as a noteworthy point in the textual chain, the analyst will not be able to detach the unit examined from its immediate context, in other words from the textual section considered pertinent for the analysis and which consists of a certain number of other linguistic events situated before and after it. Thus, when we are interested in – as is the case here – series of verb forms in the story and, more specifically, in sequences of verb tenses, we can easily understand that a form of the perfect must be studied according to its temporal frame: whether this perfect is exclusively preceded and followed by other perfects or if on the contrary it springs from a context where perfects, imperfects and pluperfects alternate, its contribution to the textual dynamic and structuring will doubtless not be the same. It is thus necessary to equip ourselves with the methods to evaluate the homogenous or heterogeneous quality of this context.

The method, borrowed from topology, is the following: we work on a textual chain reduced to the succession of grammatical labels provided by the lemmatiser and morpho-syntactical analyzer; we further reduce this chain to just the verb tense codes, which gives us a series of bicodes; for example, for the beginning of *Civ2* we obtain the following:[9]

11 12 11 11 11 11 11 11 12 12 12 12 12 12 12 12 12 12 11 11 11 11 11 15 15 12 15 15 **14** 11 11 15 11 11 11 11 11 11 11 15 15 15 11 12 12 15 **14** 12 12 12 12 12 12 **14 14 14 14** 11 **14** etc.

For each occurrence of a verbal code, we determine a contextual span of an arbitrary size[10] which constitutes the code's neighbourhood: the neighbourhood is here a size 11, in other words of 5 codes before and 5 codes after the pivot-occurrence. We then allocate to this neighbourhood a measurement that is calculated according to the property judged relevant, in this case the number of occurrences (the "density") of the perfect indicative (indeed, the perfect is considered the basic tense of the narrative discourse). To do this, we simply count the number of codes "14" present within the textual span identified as a neighbourhood (the

[8] This definition does not exclude the possibility of also considering the text in its 'reticular' dimension, in other words as constituted of tangled thematic networks of which some are purely internal and others echo anterior texts. On this point see (Viprey, 2000 and 2002).

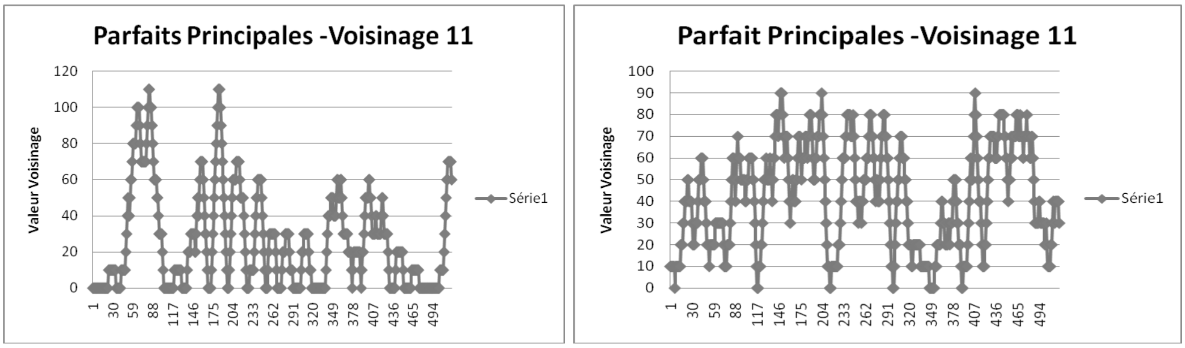
[9] Code 11 represents the present indicative, code 12 the imperfect indicative, code 14 the perfect indicative and code 15 the pluperfect indicative.

[10] In reality the optimal size of this neighbourhood has been determined by a series of tests; we can in addition work with several sizes and thus construct a family of neighbourhoods.

measurement varies between 0 and 11). This provides a discrete topological representation of the text which makes apparent its narrative dynamic in light of the use of perfect indicative[11] and which can be imaged by a simple Excel curve. The same operation can be carried out for each of the four parameters looked at in this study. By superimposing the curves, the significant joining zones of these parameters will appear and point the way towards an automatic recognition of motives.

3.2 Results and comments

Here are the results[12] obtained for *Civ 2* and *Ann 12*:



Figures 1 and 2: The textual dynamic of perfects in *Civ 2* and in *Ann12*

In terms of comparison it can be clearly seen that, on the basis of the two curves, the two dynamics are appreciably different.

In terms of the narrative's internal structure, we would be tempted to consider that a certain type of episode begins around sentence 40 and comes to an end around sentence 100 of the *Civil Wars*, and that another type runs from sentence 436 to 494 of the same book. The same is the case between sentences 320 and 360 of book 12 of the *Annals*. Nevertheless, only going back to the texts can confirm this and the signs remain weak.

On the other hand, if we superimpose the curves which represent, through exactly the same method, the dynamic of the perfects, that of the presents, that of the framing motif sentences and that of the sentences with an appended element, we obtain a cluster of extremely interesting signs, the interpretation of which becomes easier in terms of textual organization. The question of the independence of the chosen parameters from one another cannot go unaddressed. While the relative distributions of presents and perfects cannot be independent from one another, they are not *a priori* complementary: in descriptive passages, perfect is also in competition with imperfect, or, in fast narrations, with narrative infinitives. Accordingly, present and perfect complementary curves would be very instructive. Further, the relationship between framing motives and appendage motives within sentences is even freer: Latin sentences can indeed contain both motives at the same time or neither of them. Finally, there is no *a priori* syntactic or stylistic correlation between sentence structure and the tenses of the

[11] The window being a sliding one, the method allows this dynamic to be understood in its continuity. In addition, even if both methods share some similarities, our method differs from the n-grams method by being purely descriptive and in no way inferential.

[12] Distribution of Main Clause Perfects – Neighbourhood 11. On the Y-axis, the values (whose maximum is 11) have been multiplied by 10 in order to facilitate reading of the graph.

main verbs. In addition, we only superimpose distribution curves in order to reveal text structure indicators. We do not use statistical correlation tests, which require more caution with respect to the independence of the observed variables.

Below, we separately present the results for each text (*Civil War 2* and *Annals 12*) and then offer some synthetic conclusions pertaining to both works.

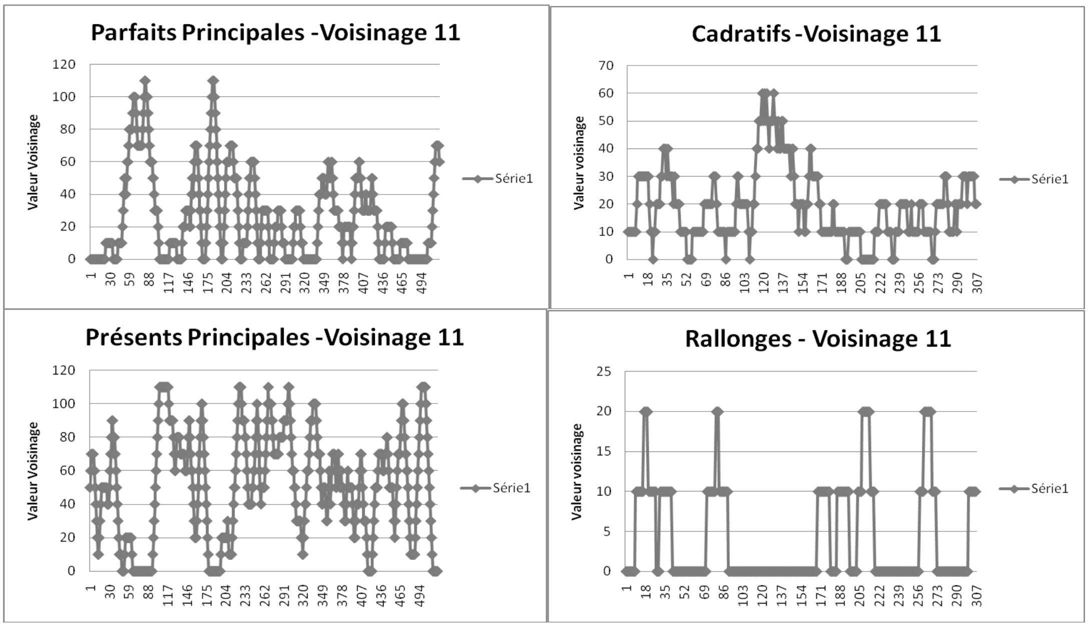


Figure 3: Distribution of the four text structure indicators in *Civ2*

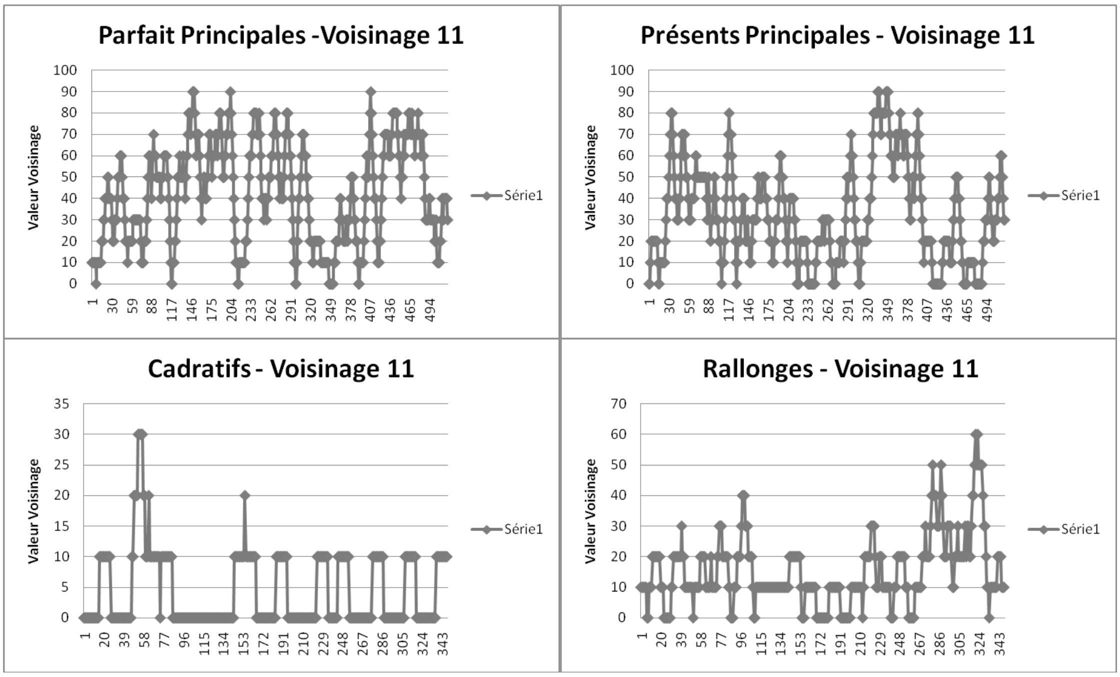


Figure 4: Distribution of the four text structure indicators in *Ann12*

A detailed examination of the curves, together with a return to the texts, allows the following conclusions to become clear:

- For both authors, the distribution of the perfect and the present is largely complementary: as previously noted, this result demonstrates that the two verb tenses offer the writers two resources with free variation; their use is thus generally guided by the writing choices linked to the principles of textual organisation.
- The framing motif sentences are less numerous in Tacitus: this is a well known fact. Several peaks in the curves are nevertheless observed: the largest of them is found between sentences 45 and 60 (§12 à 15), in which we can even detect some short typical narrative sentences of Caesarean inspiration. This passage is devoted to the Empire's external affairs (managing a crisis in Armenia; military account). Appendages are to be found there, but they are a lot rarer and shorter than in the rest of the work. We can also observe a small peak around sentence 155 (§35), which corresponds to the end of an episode (beginning around §31) recounting, in a particularly rapid rhythm, a decisive military action (the propraetor Ostorius defeats Caracatus).
- On the other hand, as expected, there are a great many framing motif sentences in Caesar's text, especially in the central narrative passage. There is, however, a "gap" between sentences 205-220, due to a direct reported speech (Curion's speech § 32 *et sq*). In this passage, we find of course a peak for the present of the indicative. Elsewhere, the small number of framing motives often corresponds to passages where the historical present is used: this tense is typical of light and fast narration, in passages free of additional reflections and of heavy descriptions of details and circumstances.
- As for the appendages, they are rare and seem erratic in *Civ2*. On the contrary, they are a lot more numerous in *Ann12*: they are one of Tacitus' main stylistic markers. The greatest density is observed before the final episode of the death of Claude, in other words in a passage devoted to internal affairs. This episode is accompanied with a slowing down of the narrative and philosophical and/or political reflections. When the narrator arrives at the death of Claude properly speaking, the appendages disappear almost entirely.
- Generally the appendages peaks coincide rather with a narrative written in the perfect tense: we are thus dealing with the basic configuration of Tacitus' narrative, in which the recounting of events is enriched by philosophical and/or political reflections. The marked rise in historical presents, markers of an acceleration of the story telling, goes hand in hand with a drop in the number of appendages; some episodes relating a strictly military operation reactivate certain Caesarian stylistics, and in particular the conjunction of a framing motif sentence with a series of historical presents. The superimposition of markers is accompanied with an alternation of episodes devoted to internal affairs and episodes devoted to external affairs and which characterizes the organisation of this text, and stresses the importance of these episodes.

The method thus has a certain interest: it brings into view sections and parts, which can be confirmed by going back to the text, even when we are faced with literary texts whose structural organisation is less flaunted than that of technical texts.

Nevertheless this method is only effective for phenomena which are sufficiently represented throughout the text: for example, it works badly for appendages in Caesar, which are extremely rare. And yet, because of the very same rarity, we could consider that the appearance of an appendage in this author is significant and maybe also functions as a textual signal.

The neighbourhood method has another disadvantage: it is difficult to evaluate on a curve if the intermediate peaks and troughs are really significant or not.

That is why we propose to round out this preliminary analysis by making use of a mathematical tool allowing for a more precise evaluation of the significance of the distributional facts observed: this tool is the bursts method (fr.: “méthode des rafales”) of Pierre Lafon (Lafon, 1981).

4 The bursts

4.1 A presentation of the method

The method aims to provide a statistical evaluation based on the ‘configuration given shape by the occurrences of a form in an ordered sequence of a text’ (Lafon, 1981: 158), in terms of the distribution’s regularity or irregularity. Let us take a simple example: if a text consists of a 1000 words and 10 occurrences of word M, a distribution of one M every 100 words would be remarkably regular; a distribution which on the other hand brought together 8 out of the 10 occurrences between the 100th and 115th word would be a striking grouping, what Lafon calls a ‘burst’; anything between the two would be an unremarkable distribution which would correspond to a random distribution.

We apply this evaluation method to the sentences with an appended element in Book 2 of the *Civil Wars*. The data necessary for such a calculation are thus: the total number of sentences (T), the number of sentences with an appended element (f), and the numerical order of these sentences with an appended element so as to calculate their various spacings (x_i). On the basis of that, we calculate an index X and we compare its value to its mathematical expectation Exp(X) [13].

4.2 Application to the appendage parameter in Civ2

With the values T = 307, f = 12, X = 1663 and Exp(X) = 556 the results of the calculation are clearly significant: the index is a lot higher in number than the expected average value and a configuration with ‘bursts’ clearly comes to the fore. A return to the text allows us to see that this distribution has significance: it always corresponds to moments when Caesar’s army is in difficulty and/or the military action is making no headway: putting in place the siege of a town, the setting up of camp on unfavourable terrain; the appendages are there to explain and justify the slowing down of the action, and maybe to give a mimetic narrative image of the events.

5 Conclusion

We hope that we have demonstrated the benefits of simultaneously taking into account several textual organisation signals in order to foreground the narrative structuring of our

[13] Exp(X) = (T-f)*(T-f-1)/f*(f+1). Exp(X) corresponds to a random spacing. The average X = (x_i*(x_i – 1))/2 is compared to Exp(X). If this average is lower than Exp(X) – a standard deviation, then the distribution is highly regular; if higher than Exp(X) + a standard deviation, then the distribution is highly irregular (*i.e.* in burst); between these two values, the distribution is a random distribution about which we cannot say anything.

corpus. In particular the stylistic characteristics (a framing one for Caesar, appendages for Tacitus) draw attention to themselves as textual organisation indicators by means of their insertion within a larger motif and must be analysed in a multi-dimensional perspective which demands the different levels of textual annotation. Therefore, to fully comprehend the text structure, we must take into account not only these convergences between indicators, but also the repetition of the convergences.

We have also demonstrated that there exist, within the framework of corpus linguistics, tools and methods to bring to light the macro-textual structuring schemas through a semi-automatic process and in leaning on statistical ratification. But these tools in no way do away with going back to the text: such a return remains necessary in order to interpret the distributional dynamics (above all in a literary context) and to better understand the articulation between the narrative’s macro-structural organisation and its micro-structural constraints. From a cognitive point of view, we feel justified in supposing that listening to or reading the text recalls a memory of these complex linguistic structures; in this way, as the listening or reading progresses, the structures gain, in a probably cumulative process, the status of fully relevant textual organisation signals.

References

ADEMA S. (2007). « Discourse modes and bases in Vergil’s *Aeneid* ». *In* R.J. ALLAN & M. BUIJS (eds) *The Language of Literature. Linguistic, Approaches to Classical Texts*, (Amsterdam Studies in Classical Philology XIII). Leiden: Brill, 42-64.

ADEMA S. (2008). *Discourse Modes and Bases. A Study of the Use of Tenses inVergil’s Aeneid*. Diss. VU University Amsterdam.

ADEMA S. (2009). « Discourse Modes and Bases. The Use of Tenses in Vergils’s *Aeneid* and Livy’s *Ab Urbe Condita* ». *Belgian Journal of Linguistics* 23 (“*New Approaches in Textual Linguistics*”), 139-152.

BURROW J. (1987). « Word-Patterns and Story-shapes: the Statistical Analysis of Narrative Style ». *Literary and Linguistic Computing*, 2/2, 61-70.

CHAROLLES M. & PRÉVOST S. (éds) (2003). *Travaux de Linguistique* 47 : *Adverbiaux et topiques*.

CHAROLLES M. & PÉRY-WOODLEY M-P. (éds) (2005). *Langue Française* 148 : *Les adverbiaux cadratifs*.

CHAUSSERIE-LAPRÉE J.P. (1969). *L’expression narrative chez les historiens latins. Histoire d’un style*. Paris : E. de Boccard.

DENHIÈRE G. & ROSSI J.P. (éds) (1991). *Text and Text Processing*. Elsevier Science Publishers, North-Holland.

HOLMES D.I. (1998). « The Evolution of Stylometry in Humanities Scholarship ». *Literary and Linguistic Computing* 13/3, 111-117.

KROON C. (2007). « Discourse modes and the use of tenses in Ovid's*Metamorphoses* ». *In* R.J. ALLAN & M. BUIJS (eds) *The Language of Literature. Linguistic, Approaches to Classical Texts*, (Amsterdam Studies in Classical Philology XIII). Leiden: Brill, 65-92.

KROON C. (2009). « Text Structure and Referential Choice in Narrative. The Anaphoric Use of theLatin Demonstrative *ille* ». *Belgian Journal of Linguistics* 23 ("*New Approaches in Textual Linguistics*"), 121-138.

LAFON P. (1981). « Statistique des localisations des formes d'un texte ». *Mots* 2, 157-187.

LOISEAU S. (2007). « CorpusReader : un dispositif de codage pour articuler une pluralité d'interprétations », *Corpus* 6, 153-186.

LONGRÉE D., LUONG X. & MELLET S. (2004). « Temps verbaux, axe syntagmatique, topologie textuelle : analyse d'un corpus lemmatisé ». *In* G. PURNELLE, C. FAIRON & A. DISTER (éds) *JADT 2004*, 7èmes Journées internationales d'Analyse statistique des Données Textuelles. Louvain : UCL, vol. 2, 743-752.

LONGRÉE D., LUONG X. & MELLET S. (2008). « Les motifs : un outil pour la caractérisation topologique des textes ». *In* S. HEIDEN ET B. PINCEMIN (éds) *JADT 2008*, Actes des 9èmes Journées internationales d'Analyse statistique des Données Textuelles. Lyon : Presses universitaires de Lyon, vol. 2, 733-744.

MELLET S. & LONGRÉE D. (2009). « Syntactical Motifs and Textual Structures ». *Belgian Journal of Linguistics* 23 ("*New Approaches in Textual Linguistics*"), 161-173.

PIERRARD S., DEGAND L. & BESTGEN Y. (2004). « Vers une recherche automatique des marqueurs de la segmentation des discours ». *In* G. PURNELLE *et al.* (éds), *JADT04, Le poids des mots*. Louvain : UCL, vol. 2, 859-864.

VIPREY J.-M. (2000). « Hypertexte de corpus littéraire : cartographie et statistique multidimensionnelle ». *In* M. RAJMAN & J.C. CHAPPELIER (éds) *JADT 2000*, 5èmes Journées internationales d'Analyse statistique des Données Textuelles. Lausanne: Ecole Polytechnique Fédérale de Lausanne, vol. 2, 535-539.

VIPREY J.-M. (2002), « Dynamisation de l'analyse micro-distributionnelle des corpus textuels ». *In* A. MORIN & P. SÉBILLOT (éds) *JADT 2002*, 6èmes Journées internationales d'Analyse statistique des Données Textuelles. Saint-Malo: IRISA/INRIA, vol. 2, 779-790.

Signalling Elaboration: Combining Gerund Clauses with Lexical Cues

Clémentine Adam, Marianne Vergez-Couret
University of Toulouse
clementine.adam@univ-tlse2.fr
marianne.vergez@univ-tlse2.fr

Abstract. In this paper, we aim at automatically identifying *Elaboration*. This relation is particularly difficult to spot since it does not have prototypical markers. Our approach focuses on an ambiguous syntactic pattern, the gerund clause, combined with lexical cues. This approach allows us to detect few but accurate cases of inner sentence Elaborations in our corpus, validating the fact that lexical cues are relevant for this task.

Keywords. Elaboration, lexical cues, distributional neighbourhood, SDRT, discourse analysis

1 Introduction

Description and detection of discourse structure is a major topic of ongoing research (Moore & Wiemer-Hastings, 2003; Péry-Woodley & Scott, 2006). Many formal and functional approaches attempt to model discourse through relations between segments (typically clauses) (Asher & Lascarides, 2003; Grosz & Sidner, 1986; Hobbs, 1990; Mann & Thompson, 1987; Wolf & Gibson, 2006). Anaphora resolution, temporal order of events identification and others empirical problems require knowledge of discourse structure (Grosz & Sidner, 1986; Lascarides & Asher, 1993; Hobbs, 1990). Applied approaches (Baldridge & Lascarides, 2005; Lin *et al.* , 2009; Subba & Di Eugenio, 2009) aim to handle and detect elements of this structure studied by formal and functional approaches in order to develop applications like automatic generation (McKeown, 1985) and automatic summarization (Marcu, 2000), among other natural language processing tasks.

In this paper, we focus on the *Elaboration* relation and on its automated identification, using SDRT's theoretical framework. The *Elaboration* relation is particularly difficult to spot, since it does not have a prototypical lexical marker according to Knott (1996). According to SDRT, the *Elaboration* relation can be lexically marked, but this hypothesis has not yet been tested on the basis of corpus data. We investigate this claim using lexical cues to identify *Elaboration*. This investigation is carried out in the framework of the VOILADIS project[1], which aims to demonstrate the importance of lexical cues for discourse analysis. More specifically, we discuss the lexical resource that we employ to highlight these lexical cues. A practical experiment of inner sentence *Elaboration* detection is presented, combining a lexical resource based on the

[1] financed by the Cluster for Higher Education and Research "Université de Toulouse" (PRES)

computation of distributional similarity on the one hand, and a weak marker of the *Elaboration* relation, the gerund clause, on the other hand.

Our first aim is to contribute towards the automated identification of the *Elaboration* relation. Our second aim, which follows from the first, is to improve the description and formalisation of this rarely studied relation, in order to expand studies on discourse signalling. We offer two contributions towards this goal: first, we extend the study of the devices that are used to mark this relation by showing that it is lexically signalled. Second, we collect examples that could be used to evaluate the adequacy of theoretical frameworks to real-world data.

2 On the *Elaboration* relation

2.1 *Elaboration* within the framework of SDRT

Segmented Discourse Representation Theory (Asher 1993, Asher and Lascarides 2003) is a formal theory of discourse. SDRT is an explanatory model making use of semantic information, world knowledge and pragmatic principles in order to explicit the rhetorical link between clauses. Discourse relations are described in two steps: first, they are associated with *triggering rules* to infer them and second, discourse relations entail *semantic effects*. SDRT claims that while discourse structure must be sensitive to non-linguistic information like world knowledge, it is conceptually and computationally more efficient to take into account linguistic knowledge to which we have direct access. *Triggering rules* to infer relations use both linguistic cues like discourse markers, syntactic constructions, verb tense, aspect and mood, argument structure, logical operator, quantifiers ; informations about lexical semantics ; and non-linguistic information about word knowledge and pragmatic principles.

The relation of *Elaboration* relates two propositions only if the second proposition provides more detail about the eventuality (state or event) described in the first. In the SDRT framework, the *triggering rule* to infer *Elaboration* is based on information about lexical semantics and world knowledge. More specifically, *Elaboration* can be non-monotonically inferred if there is a subsumption relation between the types of the eventualities involved. The Subtype predicate ($Subtype_D$) means that the type of the second eventuality is a subtype of the first one in the lexical semantics of the predicates or by some piece of shared knowledge dependent on the given discourse (D). For instance, *Elaboration* is inferred between constituant π_a and π_b (representing respectively segments (a) and (b)) in the following example :

(1) (a) Martha ate a lovely meal. (b) She devoured lots of salmon. (Asher & Lascarides, 2003, p.282)

We can non-monotonically infer that the type of the second event "devour lots of salmon" is a subtype of the first one "eat a lovely meal" thanks to lexical semantics.

Non-monotonically means that this inference can be cancelled if other monotonic inferences are established like in the following example :

(2) (a) Martha ate a lovely meal. (b) And then she devoured lots of salmon.

The discourse markers "And then" monotonically indicates that π_b (representing (b)) is attached to π_a (representing (a)) by *Narration.*

In the framework of SDRT, the lexicon is an important (but not exclusive) information source for inferring $Subtype_D$ predicate. The lexicon includes information about the semantic type of objects that are denotated by common nouns, verbs and so on. A subtype is related to a supertype by some notion of substituability: the subtype inherits many supertype characteristics and has some specific differences ; the subtype can be substituted by the supertype but the reverse is not necessarily true. The concept of subtype is closely related to the linguistic notion of hyperonymy.

In the example (1), sentences (a) and (b) include words that are semantically linked. First, the type of the event described in e_b "devour" is a subtype of the type of the event described in e_a "eat". Second, the word $meal$ must be lexically specified to be of type $food$ and $salmon$ is also of type $food$ but this lexical information is not directly coded in the type hierarchy. More lexical information are needed like, for instance, that the property of the event "meal" is to eat it ; that all words of type "food" have this property ; and that "salmon" is food derived from the animal salmon.

This information at the lexical level between predicates ("eat" and "devour") and arguments ("meal" and "salmon") sharing a same $\theta{-}role$ (here patient) allow us to infer $Subtype_D$ between the constituants π_a and π_b (build from a and b) at the discourse level.

2.2 Signalling of *Elaboration*

(Scott & de Souza, 1990; Knott, 1996; Knott *et al.* , 2001) observed that *Elaboration* is a relation for which there are no obvious surface signals, so that automatic identification using prototypical discourse markers is impractical. It is therefore necessary to find different ways (other than traditional markers) to automatically detect this relation. Marcu (2000) uses algorithms based on discourse markers and word co-occurences, and finds that discourse markers "specifically" signal *Elaboration.* However, that marker is not frequent and covers few cases of *Elaboration.* Marcu (2000) also reports on a non-linguistic marker, based on the number of sentences in a paragraph or the number of paragraphs in a section : If this number is small and no discourse markers are used, the relation between the sentences or paragraphs is generally *Elaboration. Elaboration* is particularly difficult to spot also because discourse markers are generally ambiguous, as shown by Bras2007 for the french adverbial "d'abord" *(first)* that requires subordination with a constituent above him in the discourse structure via *Elaboration Explanation Result* or *Flashback* relation. In this paper, we investigate the use of lexical cues to detect *Elaboration*, as suggested by the SDRT model. However, this development is not straightforward; the next section discusses difficulties inherent to the subtype predicate and the requirements for the resource used in this task.

3 Using lexical cues for identification of *Elaboration*

3.1 From $Subtype$ to lexical similarity

At first glance, it may seem that a resource providing information about hypernymy could be the right resource in order to detect automatically *Elaboration.* However *Elaboration* exhibits a wider range of lexical relations. The *Elaboration* relation, at the discourse level, is based on relations at the lexical level; however, those relations are diverse and not restricted to the lexical subtype relation. Since these relations emerge in discourse, the lexical phenomena involved can be different from these found in classical resources. Such relations can be established by discourse, and may be tightly related to a specific enunciation (Mortureux, 1993). We illustrate this issue in the following examples.

(3) [Un véhicule a effectué une spectaculaire sortie de route, hier vers 18 h 15, sur l'A36.]1 [La voiture circulait dans le sens Mulhouse-Montbéliard]2 [lorsqu'après être passée à hauteur du 35e RI,]3 [elle a quitté la chaussée sur sa droite.]4
[A vehicle left the road in a spectacular fashion yesterday around 6.15 on the A 36.]1 The car vas travelling from Mulhouse to Montbéliard]2 [when after reaching the 35th RI,]3 [it left the road on the right-hand side.]4

In example (3), three lexical links allow us to infer $Subtype_D(\pi_1, \pi_4)$: "véhicule" (*vehicle*)/"voiture" (*car*), "sortie" (*exit*)/"quitter" (*leave*) and "route" (*road*)/"chaussée" (*roadway*). While the first link, "véhicule" (*vehicle*)/"voiture" (*car*), is clearly classified as hyperonymy, the "route" (*road*)/"chaussée" (*roadway*) link is in fact meronymy, and the "sortie" (*exit*)/"quitter" (*leave*) link is more subtle to categorize, since cross-category relations are generally not listed in typologies.

(4) (...) [qui rappelle la vocation des bénévoles de l'association :]32 [être un soutien pour la paroisse,]33 [apporter une petite contribution financière aux travaux grâce aux manifestations et aux dons,]34 [accomplir de multiples tâches et démarches touchant aux bâtiment paroissiaux,]35 [contribuer à la convivialité entre les paroissiens.]36
[...which calls to mind the role of the Association's volonteers]32 [in being a support to the parish,]33 [in contributing a small amount financially to works through activities and donations,]34 [in completing many tasks and procedures dealing with the parish buildings]35 [and in contributing to parishioners' conviviality.]36

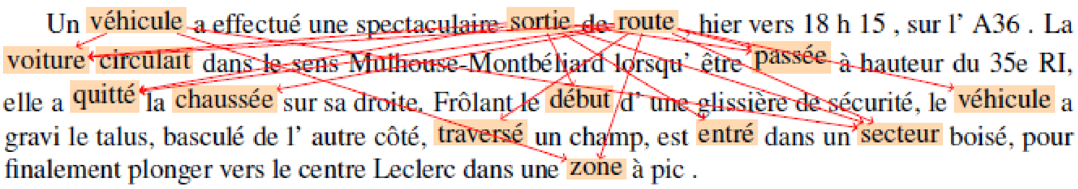
Here, events in segments 33 to 36 are subtypes of "vocation des bénévoles de l'association" (role of the Association's volonteers). At the word level, $Subtype_D(\pi_{[}32], [\pi[33] - \pi_{[}36]])$ rests on links between "vocation" (*vocation*) and words such as "soutien" (*support*), "accomplir" (*to complete*), "tâche" (*task*) or "contribuer" (*to contribute*). These links are established in discourse, and will most probably not appear in a generic resource, since they do not match a classical lexical relations. Such links are more accurately referred to as lexical similarity relations.

3.2 Selecting the appropriate lexical resource: distributional neighbours

We have seen that the *Elaboration* relation seems indeed lexically marked, but that the links involved are softer than subtype. In order to automatically detect this relation, the resource chosen is crucial: it should contain these links for their automated usage. As stressed in the previous section, a generic resource seems poorly fitted to this task. We have focused on a resource built from corpora, taking into account semantic proximity links, possibly across parts of speech. In particular, we have chosen the *Voisins de Wikipédia* database, a resource built by distributional analysis. The principle of distributional analysis is to pair words based on their shared contexts, following Harris (1968) hypothesis. The paired words share second-order affinities: they do not need to appear together in the corpus, but their environments are similar (Grefenstette, 1994). The lexical relations put in evidence are then paradigmatic.

The *Voisins de Wikipédia* database was build from a full archive of the online encyclopedia Wikipedia, which contains more than 194 millions words. The archive was processed through the Syntex-Upéry chain developed by Bourigault (2002). First, a syntax analysis is performed. Then, all <governor, relation, dependant> triplets are listed, an example triplet is: <circuler, à bord de, voiture> (<*travel, in, car*>). The triplets are then transformed in <predicate, argument> couples, where the predicate is a combination of two components: the governor and the relation, in the previous example <circuler_à bord de, voiture> (<*travel_in, car*>). The similarity between distributions is computed for each predicate couple and each argument couple using Lin's score: Predicates are paired based on their shared arguments; reciprocally, the same pairing is performed on arguments, based on their shared predicates. Thus, arguments "véhicule" (*vehicle*) and "voiture" (*car*) are paired through predicates such as "circuler_à bord de" (*travel_in*), "capot_de" (*hood_of*), "conduire_obj" (*to drive_obj*) , etc.

The obtained resource contains 4 million pairs, covering a large panel of relations. An example of neighbourhood links projected on the text sample (3) is provided below. Only links between two sentences appear.



Un véhicule a effectué une spectaculaire sortie de route hier vers 18 h 15 , sur l' A36 . La voiture circulait dans le sens Mulhouse-Montbéliard lorsqu' être passée à hauteur du 35e RI, elle a quitté la chaussée sur sa droite. Frôlant le début d' une glissière de sécurité, le véhicule a gravi le talus, basculé de l' autre côté, traversé un champ, est entré dans un secteur boisé, pour finalement plonger vers le centre Leclerc dans une zone à pic .

Here, aforementioned links relevant for identifying the *Elaboration* relation are observed: "véhicule" (*vehicle*) / "voiture" (*car*), "sortie" (*exit*) / "quitter" (*leave*), and "route" (*road*) / "chaussée" (*roadway*). Other links participating in global lexical cohesions are observed, but these links are not involved in the *Elaboration*. Finally, many links are not relevant in this context, for example "route" (*road*) and "traverser" (*to cross*).

The plethoric nature of this resource is a strong barrier against its broad usage. Even though relations relevant to our task are put in evidence by projecting neighbours in the text, many other irrelevant neighbourhood links will interfere, making a direct inference to the discourse level impossible. It is therefore necessary to define more restrictive markers, by taking into account more elaborate criteria than the simple presence of neighbourhood links. We choose to experiment on detection based on targeted neighbourhood links combined with the presence of a weak elaboration marker: gerund clauses.

3.3 Combining the neighbours with a weak cue: the gerund clause

Since the information provided by the neighbours is too plethoric we propose to combine it with an ambiguous cue of *Elaboration*. Such a combination should be more reliable than each cue considered separately since the conjunction of two ambiguous cues builds a stronger cue. Gerund clauses are the perfect candidate for this combination: some gerund clauses could be considered as elaborations of the main clause and they are easy to extract with SYNTEX.

The gerund clause establishes a syntactic subordination between two verbs: two processes are linked in this way. The different semantic values expressed by the gerund clause are not conveyed by the gerund itself but depend on the combination of the two linked verbs. The interpretation is done *a posteriori* and determined by the semantic relationship between the verbs and other elements given by the context (Halmoy, 1982).

Our analysis reveals that gerund clauses in French can be linked to the main clause with two main discourse relations. For (5) and (6), we infer an *Elaboration* relation:

(5) Les Britanniques réagissent en emprisonnant ou en tuant les derniers chefs.
The British react by imprisoning or killing the last leaders.

In (5), the main clause introduces an underspecified event "réagir" *to react* and the two gerund clauses introduce two events that specify it "emprisonner" *to imprison* and "tuer les derniers chefs" *to kill the last leaders*.

(6) Puis on irrigua les alentours en creusant un canal derivé du Zab Supérieur.
Then, the surrounding areas were irrigated by digging a canal leading from the River Zab Supérieur.

In (6), the main clause introduces the event "irriguer" (*to irrigate*) and the gerund clause introduces the event "creuser" (*to dig*). "Irriguer" and "creuser" are semantically linked. The activity denoted by the type of event "irriguer" can involve the activity denoted by the type of event "creuser". So, we can infer $Subtype_D$ between the type of events "irriguer" and "creuser".

However, gerund clauses are not always elaborations of the main clause like in the following example:

(7) Dans la ville de Koriko, Kiki, accompagnée de son chat noir Jiji, va distribuer des colis en volant sur son balai, grâce à ses pouvoirs.
In Koriko town, Kiki, with her black cat Jiji, delivers parcels while flying on her broom, thanks to her magical power.

In (7), the gerund clause gives background of the main clause. The *Background* relation are typically used for setting the stage of an event. In (7), the main clause introduces the event "distribuer des colis" (*to deliver parcels*) and the gerund clause gives background information "voler sur son balai" (*to fly on her broom*).

Our main idea is that verbs and objects in the main clause and the gerund clause will generally be neighbours in *Elaboration* cases and not in *Background* cases. Considering our examples (6) and (7), it seems to us that "irriguer" (*to irrigate*) and "creuser" (*to dig*) could be found as neighbours but not "distribuer" (*to deliver*) and "voler" (*to fly*).

With this hypothesis in mind, we set up the experimentation presented in the next section.

4 Experimental validation

4.1 Motivations and strategy

The goal of the presented experiment is to reliably identify *Elaborations*: we aim for the highest precision. This task is challenging: it is sparsely attempted in the literature and the attained reliability is low. Nevertheless, such attempts are required for a better understanding of the *Elaboration* relation.

While this task is interesting in itself, our experiment will also illustrate the improvement brought by taking into account lexical phenomena for discourse analysis, and show the relevance of lexical neighbourhood for detecting these phenomena. If using lexical neighbours brings a significant performance improvement, we will also validate the fact that *Elaboration* is a lexically marked relation.

In order to reach these goals, we chose to use the lexical neighbours in combination with a weak clue for *Elaboration*, the gerund clause. Two combination strategies are tested.

- In a first run, *Elaboration* is detected if verbs in the main clause and in the gerund clause are neighbours (these candidates are noted by GN in the following), *Cf.* (8).
- In a second run, *Elaboration* is detected if verbs in the main clause and in the gerund clause are neighbours, **and** if the verbs objects are connected by at least one neighbourhood link (GNON in the following), *Cf.* (9).

(8) ... et les villages *contribuaient* également à ce grand projet religieux *en envoyant* des vivres.
(...) and villages also contributed *to this great religious project by* sending *supplies*

(9) Les Skrulls (...) *élargissent* leur **empire** en *englobant* dans celui-ci les **mondes** moins avancés qu'ils rencontrent.
The Skrulls (...) were expanding *their* ***empire*** *by* incorporating *the less evolved* ***worlds*** *they discovered.*

We will compare the results of these two combination strategies to the results obtained by considering only the gerund clause (G in the following). The observed performance differences will allow us to quantify the improvement brought by using lexical neighbours to enhance *Elaboration* detection.

4.2 Extraction of *Elaboration* candidates

In this experiment, the corpus used is a fraction of the french Wikipedia: 45'823'899 words from 5'106'831 sentences, which amounts to roughly one fifth of the online encyclopedia. This corpus has been pre-processed with SYNTEX. All sentences featuring a [verb clause, gerund clause] pair are extracted (G). Two subsets of these candidates are then produced, by taking into account constraints of lexical neighbourhood on the verbs pairs (GN) and on the verb objects pairs (GNON), as explained in the previous section. The following table gives the number of candidates obtained depending on which markers were used.

G	GN	GNON
18571	375	193

The number of *Elaboration* candidates is small considering the corpus size. Nevertheless, in the current state of research on this relation, defining a reliable marker is a significant improvement, even if the number of matches is small.

4.3 Annotation of extracted candidates

Each text was independently annotated by two experts in discourse relations[2]. We annotated 314 examples, approximatively 100 for each case (G, GN, GNON) presented randomly to the annotators with the question: *Is the gerund clause an elaborating segment of the main clause?*

The agreement rate between experts is 89% (280 cases of agreement *vs* 34 cases of disagreement). The kappa score (Cohen, 1960) is 0.70, which highlights a moderate to good inter-annotator agreement. This reveals the difficulty of the task. The kappa score is, however, good enough to consider an automatisation of this task.

In a second run, we explored the 34 examples for which we disagreed, in order to make sure the reference annotation was as reliable as possible and to analyze the types of inter-annotator variation. The discussion allowed us to refine the annotation for the vast majority of disagreement cases. Finally, only 9 cases resulted in the experts disagreeing; such cases include texts for which two interpretations are possible. To ensure meaningful results, these 9 marginal cases were discarded from the reference which was subsequently used to evaluate the results of the automated elaboration detection, cf. next section.

4.4 Results and perspectives

The table below summarizes the results obtained when testing the three strategies for *Elaboration* detection.

	Extracted	**Annotated**	**Elab.**	**Not Elab.**	**Precision**	**Confidence interval**
G	18571	102	62	40	**60.8%**	9.45%
GN	375	100	81	19	**81.0%**	6.59%
GNON	193	104	99	5	**95.2%**	2.8%

These results confirm that the gerund clause is indeed an ambiguous cue, since only 60.8% of the candidate are *Elaboration*. The number of annotated candidates is small considering the amount of gerund clauses extracted; this results in a wide confidence interval. However, the performance difference between G and our two strategies is large enough to ensure that these two strategies bring a significant improvement. With the first strategy (gerund clause and main verb are neighbours), 81% of the cases are *Elaboration*. The second strategy (gerund clause and main verb are neighbours *and* the verbs objects are linked by neighbourhood) is very reliable, with 95% precision. These results are highly promising.

The cases where our markers failed were analysed. In a few cases, the failure is caused by an irrelevant neighbourhood link. For example, in the context of example (10), the link between "marcher" (*march*) et "incendier" (*burn*) is irrelevant. In various cases, a different marker can be observed, which could be used to cancel the *Elaboration* inference. This is illustrated in example (11), where the strong lexical marker of *Contrast* relation "mais" (*but*) appears.

(10) Ils *marchent* la campagne en *incendiant* toutes les habitations.
They marched *the countryside,* burning *down every dwelling they found.*

(11) Le roi d'Espagne lui accorda une décoration qu'il accepta, *mais* en refusant la pension qui y était attachée.

[2] The authors of this paper.

> *The king of Spain accorded him a decoration that he accepted* while *refusing the pension that was attached to it.*

These considerations suggest that our good results can still be improved upon, by taking into account other types of markers, signalling an other discourse relation on the one hand, and by a more elaborate filtering for the neighbours on the other hand.

5 Conclusion and outlook

We have presented a practical experiment dedicated to the detection of *Elaboration*. While *Elaboration* is often considered as a relation without prototypical lexical discourse markers, our aim was to find signalling devices for the identification of *Elaboration*. We combined an ambiguous cue, the gerund clause and information provided by the lexical neighbours resource.

The results of our experiment are encouraging. We validate on a corpus the fact that *Elaboration* can be lexically marked as suggested in the SDRT framework. With this contribution, we also follow the objectives of ANNODIS project[3], which aim to construct an annotated corpus for the study of discourse organisation in order to improve description and formalisation of discourse relations with real-world data.

The prevalence of lexical cues for discourse structuration is commonly accepted, but they are still neglected in NLP applications because of the difficulty to pick out lexical links in texts. This contribution validates lexical neighbours as a relevant resource to use for this task, in the case of *Elaboration* detection. In the course of the VOILADIS project, we hope to generalize the usage of lexical cues for all aspects of discourse analysis.

Nevertheless, our approach detects a few cases of *Elaboration* in the whole corpus. Improvements could be made by detecting *Elaborations* between sentences. First we will continue to combine neighbours and weak cues of *Elaboration* such as the adverbial expressions "d'abord" (*first*), "dans un premier temps" (*at first*). Second, we will investigate the role of the neighbours by taking into account the density of neighbours between two sentences, the syntactic position of the neighbours, etc.

Acknowledgements

We would like to thank Cécile Fabre and Philippe Muller for discussions related to the content of this paper and for the help with english writing and presentation.

References

Asher, Nicholas, & Lascarides, Alex. 2003. *Logics of conversation.* Cambridge:CUP.

Baldridge, Jason, & Lascarides, Alex. 2005. Probabilistic Head-Driven Parsing for Discourse Structure. *Pages 96–103 of: Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005).* Ann Arbor, Michigan: Association for Computational Linguistics.

[3] financed by the French National Research Agency (ANR)

Bourigault, Didier. 2002. Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *Pages 75–84 of: Actes de la 9ème conférence sur le Traitement Automatique de la Langue Naturelle*.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

Grefenstette, Gregory. 1994. Corpus-Derived First, Second and Third-Order Word Affinities. *Pages 279–290 of: Proceedings of Euralex*.

Grosz, Barbara. J., & Sidner, Candace. L. 1986. Attention, intentions and the structure of discourse. *Computational Linguistics*, **12**(3), 175–204.

Halmoy, Jane-Odile. 1982. *Le gérondif. Eléments pour une description syntaxique et sémantique*. Ph.D. thesis, University of Trondheim.

Harris, Zellig. 1968. *Mathematical Structures of Language*. New-York: John Wiley & Sons.

Hobbs, Jerry R. 1990. *Literature and cognition*. CSLI Lecture Notes. version papier. Chap. 5. The coherence and structure of discourse, pages 83–114.

Knott, Alistair. 1996. *A data-driven methodology for motivate a set of coherence relations*. Ph.D. thesis, University of Edinburgh.

Knott, Alistair, Oberlander, John., O'Donnell, Michael, & Mellish, Chris. 2001. Beyond elaboration : the interaction of relations and focus in coherent text. *Pages 181–196 of:* Sanders, T., Schilperoord, J., & Spooren, W. (eds), *Text representation : linguistic and psycholinguistic aspects*. Amsterdam : Benjamins.

Lascarides, Alex, & Asher, Nicholas. 1993. Temporal interpretation, discourse, relations and commonsense entailment. *Linguistics and Philosophy*, **6**(5), 437–493.

Lin, Ziheng, Kan, Min-Yen, & Ng, Hwee Tou. 2009. Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. *Pages 343–351 of: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics.

Mann, William C., & Thompson, Sandra A. 1987. *Rhetorical Structure Theory : a theory of text organisation*. Tech. rept. Technical report ISI/RS-87-190, Information Sciences Intitute.

Marcu, D. 2000. The rhetorical parsing of unrestricted texts : a surface-based approach. *Computational Linguistics*, **26**(3), 395–448.

McKeown, K. R. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge: Cambridge University Press.

Moore, Johanna, & Wiemer-Hastings, Peter. 2003. Discourse in Computational Linguistics and Artificial Intelligence. *Pages 439–486 of:* Graesser, A., Gernsbacher, M., & Goldman, S. (eds), *Handbook of Discourse Processes*. Mahwah, NJ: Erlbaum.

Mortureux, Marie-Françoise. 1993. Paradigmes désignationnels. *Semen*, **8**, 123–141.

Péry-Woodley, Marie-Paule, & Scott, Donia. 2006. Computational Approaches to Discourse and Document Processing. *T.A.L*, **47**(2), 7–19.

Scott, Donia, & de Souza, Clarisse Sieckenius. 1990. Getting the message across in RST-based text generation. *Pages 47–73 of:* Dale, R., Mellish, C., & Zock, M. (eds), *Current Research in Natural Language Generation*. Academic Press, London.

Subba, Rajen, & Di Eugenio, Barbara. 2009. An effective Discourse Parser that uses Rich Linguistic Information. *Pages 566–574 of: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Boulder, Colorado: Association for Computational Linguistics.

Wolf, Florian, & Gibson, Edward. 2006. *Coherence in Natural Language*. The MIT Press.

Signalling genre through Theme: The case of news reports and commentaries

Julia Lavid, Jorge Arús and Lara Moratón

Dpt. English Philology I
Faculty of Philology, Universidad Complutense de Madrid
{lavid, jarus}@filol.ucm.es

Abstract The purpose of this paper is to analyse how the clausal thematic features observed in two newspaper genres –news reports and commentaries- can be interpreted as textual signals of their different generic characterisation. This is done through the qualitative and quantitative analysis of a sample consisting of thirty three English texts, divided into two groups of seventeen news reports and sixteen commentaries, respectively. The analysis focused on the following thematic features: (1) the experiential elements selected as Thematic Heads; (2) the semantic nature of the nominal elements realising these Heads and their internal structure; (3) the textual and interpersonal thematic choices as part of a multiple theme. The analysis revealed that each newspaper genre prefers certain thematic features and that the differences between both genres are statistically significant. It is suggested that these thematic preferences can be attributed to genre-related variables such as the communicative purpose or the subject matter of the text.

Keywords: newspaper genres, thematic features, reports, commentaries

1 Introduction

Within the now consolidated research strand of media discourse analysis, numerous studies have focused on certain linguistic features of the different newspaper genres, such as news stories, reports, or editorials.

Though news stories and reports have been extensively studied from different perspectives (Van Dijk 1988, Bell 1991, Fairclough 1995, White 1998, Scollon 1998, *inter alia*), newspaper commentaries have received less theoretical and empirical attention, though contrastive work on commentaries (Wang 2008), opinion columns (Dafouz 2008) and editorials (Alonso Belmonte coord. 2007, Lavid et al. 2009, Tirkkonen-Condit 1996, *inter alia*) is offering interesting results for the description of opinion newspaper genres across languages and cultures.

This study attempts to advance knowledge in this area by focusing on the clausal thematic features of news reports and commentaries as textual signals of their different generic characterization. With our analysis we hope to contribute to a profitable line of research which investigates the correlations between genre and thematic content (see Lavid, 2000; Eiler, 1986; Francis, 1989, 1990; Fries and Francis, 1992; Nwogu and Bloor, 1991, among others).

2 Defining the two newspaper genres

Genre refers to language use in a conventionalised communicative setting in order to give expression to a specific set of communicative goals of a disciplinary or social institution, which give rise to stable structural forms by imposing constraints on the use of lexico-grammatical as well as discoursal resources (Bhatia 2002). Studies of genre aim to capture how writers achieve their communicative purposes by using various structural forms, constructing different focuses and manipulating topics and readers by using various linguistic devices.

According to White (1998:243), news reports are ‘grounded in communicative events such as speeches, interviews and press releases’ and ‘act primarily to represent, not activity sequences, but the points of view of various external sources.’ They are classified as communicatively-based rather than event based. A news report should strive to remain objective and should use neutral language while presenting a diversity of opinions, voices, and perspectives of the event, incident, or issue under discussion.

News commentaries are opinion articles with the important communicative function of contributing to the formulation of certain, ‘preferred’ viewpoints about the world. The function of news commentaries within the larger context of newspaper coverage is to offer newspaper readers a distinctive and sometimes authoritative voice that speaks to the public directly about matters of public importance. Usually written by academics, journalists and other experienced native language writers, they exert an important influence on political opinion-formation, both on the everyday reader and on the institutional and/or elite members of a society.

According to the definitions provided above, the communicative purposes of news reports and commentaries differ in several respects. While the main communicative purpose of news reports is basically informative, the goal of news commentaries is analytical, evaluative and persuasive. In functional theories of language, differences in contextual choices – such as the communicative purpose of a given genre- are reflected in the linguistic features used by language producers. It should be expected, therefore, that the different communicative purposes which characterise these two newspaper genres would be reflected in their thematic features. This is the basis for the research hypothesis presented below.

3 Research questions

Following the line of reasoning outlined above, this study examines the clausal thematic features used by writers when constructing their messages in news reports and commentaries. More specifically, the present study proposes the following research questions:

1. What are the clausal thematic features selected by news reports and commentaries? Are there any differences in this selection?

2. If there are differences, can these be interpreted as textual signals of the different genres to which these texts belong?

These research questions are investigated through the empirical analysis of a sample of news reports and commentaries, as explained below.

4 Data and procedure

4.1 Data

The data used for this study is a sample of a total 901 clause complexes (all of them declarative with the exception of two interrogatives and four imperatives) belonging to two groups of texts, all of them collected from published sources between 2008 and 2009. One first group consists of seventeen newspaper commentaries written by expert writers or journalists extracted from the Project Syndicate, an international association of quality newspapers that publishes commentaries by prominent figures to the world's foremost newspapers on topics ranging from economics, political and international affairs to science and philosophy (see http://www.project-syndicate.org). The second group consists of sixteen news reports extracted from the news section of the online version of the *Times* (http://www.timesonline.co.uk) newspaper. Table 1 below lists the texts and provides information on authors, number of clauses and number of words, with the text reference to be used for examples in the rest of this paper in the first column.

Text reference	**Year**	**Author**	**Title**	**Number of Clauses**	**Number of Words**
Reports 1	2008	Adam Sage	Dominique Strauss-Kahn, head of the IMF, escapes dismissal over affair	11	341
Reports 2	2008	Carl Mortished	Opec hawks want to cut oil production to keep up price	16	422
Reports 3	2008	Suzy Jagger and Mike Harvey	“Microsoft results offer hope to tech sector”	13	315
Reports 4	2008	Lilly Peel	Dutch Government to inject €10bn of fresh capital into ING savings bank	15	347
Reports 5	2008	Leo Lewis,	G7 'preparing to drive down the yen'	15	566
Reports 6	2008	Lilly Peel	Ukraine agrees terms of $16.5bn rescue by IMF	5	121
Reports 7	2008	Grainne Gilmore	Barclays rejects government funding, secures £5.8bn from Qatar and Abu Dhabi	21	593
Reports 8	2008	Gary Duncan	US economy officially on brink of recession	18	606
Reports 9	2010	Tony Allen-Mills	Man found alive in Haiti after being buried for 11 days	22	508
Reports 10	2010	Maurice Chittenden and Chris Hastings	Haiti earthquake concert raises £35m	38	790
Reports 11	2010	James Bone	'Bin Laden' claims Christmas Day bomb plot	22	773
Reports 12	2010	Marie Woolf	All service veterans to get right to jump NHS queues	22	553
Reports 13	2010	David Leppard	Indian hijack plot caused new UK terror alert	25	569
Reports 14	2010	Sean O'Neill	Terrorist threat level raised to 'severe'	14	448
Reports 15	2010	Richard	Tony Blair faces legality blow over Iraq	32	884

Text reference	Year	Author	Title	Number of Clauses	Number of Words
		Woods and Michael Smith	war		
Reports 16	2010	Christine Seib	Obama 'confident' Senate will approve second term at Fed for Ben Bernanke	16	422
Reports 17	2010	Anne Barrowclough	Venezuelan cable television channel taken off air	19	485
Subtotal Reports				325	8743
Comment. 1	2009	Shlomo Ben-Ami	The Bigger Issue in Sudan	32	798
Comment. 2	2009	Aleksander Kwasniewski, Tadeusz Mazowiecki and Lech Walesa	The Vanishing Bomb	36	917
Comment. 3	2009	Vaclav Smil	The Limits of Energy Innovation	32	924
Comment. 4	2009	Leif Pagrotsky	Micro-Europe	49	969
Comment. 5	2009	Marcel de Haas	Central Asia’s Waking Giant	38	936
Comment. 6	2009	Martin Feldstein	The Case for Fiscal Stimulus	36	801
Comment. 7	2009	Raghuram Rajan	The Global Economy’s Dialogue of the Deaf	29	864
Comment. 8	2009	Dominique Strauss-Kahn	The New IMF	37	768
Comment. 9	2010	Erik Berglof and Kat harina Pistor	European Financial Regulation’s Wrong Turn	22	866
Comment. 10	2010	Peer Steinbrück	The Case for a Global Financial-Transaction Tax	41	876
Comment. 11	2010	Sebastian Dullien and Dan iela Schwarzer	An External Stability Pact for Europe	34	855
Comment. 12	2010	Otmar Issing	Ban the Common Bond	27	810
Comment. 13	2010	Giles Merritt	Where’s Europe?	30	777
Comment. 14	2010	Charles Wyplosz	Beggar-Thy-Neighbor Exchange Rates?	43	815
Comment. 15	2010	Charlie McCreevy	Insecuritization	34	768
Comment. 16	2010	George Soros	Time to Fix the Euro	39	839
Subtotal Commentaries				**576**	**13583**
Total				**901**	**22326**

Table 1: Information on the textual sample used in this study

4.2 Procedure

Our analysis applied the model proposed in Lavid et al. (2010) for the study of the message structure of Spanish, but extending it to deal with the typological features of English. The clausal features selected for analysis capture the three main types of meaning represented by the category of Theme in the linguistic literature: experiential, interpersonal and textual. For the analysis of the experiential meaning, we focused on the category of the Thematic Head, since this captures the nuclear experiential choice within the clause and is more central for the text’s thematic development. We also inspected the semantic nature of the nominal elements functioning as Thematic Heads and their internal structure. To complete the metafunctional analysis, we also included interpersonal and textual Themes as part of a multiple Theme in our analysis. The procedure was as follows:itations:

1. We segmented the texts into clause complexes, consisting of a main clause preceded or followed by one or more subordinated clauses.

2. We assigned the label of "Thematic field" (TF) to the material from the beginning of the clause complex up to and including the first nuclear experiential constituent, and divided this material into Outer Thematic Field (OTF) and Inner Thematic Field (ITF), respectively. The ITF consists of a nuclear Thematic Head –underlined in all the examples below- and any possible PreHead material preceding it.

3. We searched for the Thematic Head in each main clause. This is defined as the first nuclear experiential constituent within the main clause which is more central to the unfolding of the text by allowing the tracking of the discourse participants (see Lavid et al. 2010). In English the Thematic Head usually conflates with the Subject or the Complement in preverbal position of the main clause, as in example (1) below (underlined):

(1) The commitment of Sudan's government to the CPA has always been equivocal. (Comment 1)

4. We inspected the experiential roles (e.g.: Actor, Goal, Sayer, Beneficiary/Recipient, Senser, Phenomenon, Carrier, Token and Existent), selected as Thematic Heads in main clauses and annotated their frequencies. Examples (2), (3) and (4) below illustrate several cases of Thematic Heads functioning as Carrier, Actor, and Sayer.

(2) The negative stigma attached to IMF financing is a thing of the past. (Comment. 8)

(3) A high-powered Russian delegation recently arrived in Juba, the South Sudanese capital, with the proclaimed aim of "playing a more active role on the African continent." (Comment. 1)

(4) Shakour Shaalan, the fund's executive director, said 1MF staff, and notably female staff, "are not at all happy" with Mr Strauss-Kahn, whose amorous adventures in France have earned him the epithet Ie grand seducteur. (Report 1)

5. We looked at the nominal elements realizing the Thematic Heads and annotated whether they were concrete or abstract nouns. Concrete nouns refer to human participants (e.g., Mr. Tiltman); titles (e.g., the managing director); pronouns (e.g., she, he, they), groups of people or institutions (e.g., the Government, Microsoft, etc..). Abstract nouns refer to qualities or mental concepts (e.g., expectations, results, etc…).

6. We examined the internal complexity of the nominal elements and annotated their frequencies. We annotated as complex those Nominal Groups with long, varied and/or multiple Heads or Modifiers. Example (5) illustrates a complex Nominal Group functioning as Thematic Head. It consists of a determiner ('the'), a head ('agreement'), and two Postmodifiers ('by the American and Russian presidents' and 'to renew strategic arms reductions'):

(5) The agreement by the American and Russian presidents to renew strategic arms reductions has revived hope for the global abolition of nuclear arms. (Comment. 3)

7. For the analysis of the interpersonal Themes, we inspected and annotated the frequencies of the elements in clause-initial position immediately preceding or

following the Thematic Head and expressing the attitude and the evaluation of the speaker with respect to his/her message, including those expressing modality and polarity. Examples (6) and (7) below illustrate this type of Theme (in italics):

(6) *Unfortunately*, the discussion between countries on trade nowadays is very much a dialogue of the deaf, with countries spouting platitudes at one another, but no enforceable and verifiable commitments agreed upon. (Comment. 7)

(7) *Perhaps inevitably*, the SCO – and Russia and China as its leading members – regards NATO's increased presence in the region with some mistrust. (Comment. 5)

8. Similarly, for the analysis of the textual Themes, we inspected and annotated the frequencies of the elements in clause-initial position which are instrumental in the creation of the logical connections in the texts, such as linkers, binders, and other textual markers. Some examples of textual Themes are the elements "first", "second" and "third" –in bold- in example (8) below:

(8) **First**, the EU still spends far more of its resources subsidizing declining sectors than preparing for the future. […] **Second**, Europe has failed to introduce an independent European Research Council to ensure that funding is allocated on the basis of scientific merit. […] **Third**, European resources are fragmented, and this hampers European competitiveness.

5 Results

In the following subsections we will comment on the analysis results of the different clausal thematic choices, beginning with the results of the experiential roles selected as Thematic Heads.

5.1 Experiential roles as Thematic Heads

When comparing the types of experiential role selected as Thematic Head we found interesting commonalities and differences between the two newspaper genres under study. Thus, the experiential role of *Actor* presents very high frequencies in both genres, with no statistically-significant differences between them (28.31% in news reports vs. 28.12% in commentaries), as shown in tables 2 and 3 below. This is probably due to the fact that both genres deal with current events, which accounts for the high percentage of Material processes in both groups of texts.

Experiential roles	**Absolute frequency**	**Relative frequency (%)**
Sayer	104	32.00
Actor	92	28.31
Carrier	45	13.85
Goal	20	6.15
There	20	6.15
Senser	10	3.08
Process	8	2.46
Token	8	2.46
Value	6	1.85
Attributor	4	1.23
Receiver	3	0.92
Scope	2	0.62
Verbiage	2	0.62
Assigner	1	0.31

Phenomenon	1	0.31
Totals	325	100

Table 2: Experiential roles as Thematic Heads in news reports

Experiential roles	**Total frequency**	**Relative frequency (%)**
Actor	167	28.99
Carrier	162	28.12
Token	72	12.50
Goal	48	8.33
Senser	28	4.86
Sayer	22	3.82
Value	21	3.65
Process	16	2.78
Attribute	9	1.56
Attributor	9	1.56
Phenomenon	6	1.04
Scope	3	0.52
Verbiage	3	0.52
There	3	0.52
Recipient	2	0.35
Inducer	2	0.35
Assigner	2	0.35
Receiver	1	0.17
Total	576	100

Table 3: Experiential roles as Thematic Heads in commentaries

However, when examining other experiential roles selected as Thematic Heads, we can observe certain interesting differences:

- *Sayers* are much more frequently selected as Thematic Heads in news reports (32%) than in commentaries (3,82%), and this difference is statistically-significant (p<0.05). This can be explained by the high number of citations used in news reports with the purpose of presenting a diversity of opinions, voices and perspectives on the event, incident or issue under discussion, as illustrated in example (9) below:

(9) Mr Tilmant said the bank had the trust of its customers and had not seen a large outflow of funds (Report 7)

- *Carriers* are more frequently selected in commentaries (28,12%) than in news reports (13,85 %), and this difference is statistically-significant (p<0.05). This can be interpreted as a textual device used writers to give their opinion in what seems to be an objective and impartial way. By using relational structures with nominalizations to represent processes, writers of commentaries become invisible and manage to detach themselves from their opinions. For example, while a personally visible writer would use phrases like ‘I propose…’, the invisible writer would prefer ‘It is clear that…’ (Davies 1989). Examples (10) and (11) below illustrate this use:

(10) The EU’s goal of spending 1% of GDP on research by 2010 will not be met by any member state.

(11) The main expectations are for a reduction of nuclear armaments.

The rest of the experiential roles present low frequencies in both genres.

5.2 Nouns as Thematic Heads

With respect to the type of Nouns realizing Thematic Heads, we found statistically-significant differences (P<0.05) between both genres: while news reports prefer concrete nouns as realization of the Thematic Heads (63,69%) rather than abstract ones (14,46%), commentaries present the opposite proportionalities, favouring the use of abstract nouns (40,10%) over concrete ones (30,90%).

Concrete nouns in our sample refer to human participants (e.g., Mr. Tiltman, …); titles (e.g., the managing director); pronouns (e.g., she, he, they), groups of people or institutions (e.g., the Government, Microsoft, etc..). Abstract nouns refer to qualities or mental concepts (e.g., expectations, results, etc…). Arguably, news reports are more concerned with individuals or groups of people, while commentaries are more concerned with the exposition and evaluation of ideas.

5.3 Complexity of nominals as Thematic Heads

As to the internal structure of the nominals realising Thematic Heads, both genres present interesting differences. In commentaries they tend to be longer and more complex in their internal structure (17,20%) than in news reports (5,36%), and this difference is statistically-significant. Example (12) illustrates the use of a long Nominal Group as Thematic Head:

(12) As a result, its ability to maintain services – and the military capacity to respond to any maneuver by the Khartoum government aimed against the peace agreement – is seriously compromised. (Comment. 1)

Examples (13) and (14) also illustrate two complex nominals as Thematic Heads, with two long Qualifiers introduced by a Past Participle form (‘issued…’, ‘elaborated…’):

(13) The most immediate result of the arrest warrant issued for Sudanese President Omar Hassan al-Bashir by the International Criminal Court last month was the expulsion of most aid agencies from the country. (Comment. 1)

(14) The Group’s report, Arms Control Revisited: Non-proliferation and Denuclearization , elaborated under the chairmanship of Adam D. Rotfeld of Poland and drafted by the British scholar Ian Anthony of SIPRI, was based on contributions by security analysts from nuclear powers and Poland, as well as from countries previously in possession of nuclear weapons (South Africa) and post-Soviet countries where they were once stored (Belarus, Kazakhstan, an Ukraine). (Comment. 2)

By contrast, the internal structure of Nominal Groups as Thematic Heads in news reports is minimal, in comparison with commentaries. The only long Nominal Groups in news reports are structures consisting of Noun + Apposition whose purpose is to uniquely identify the discourse participant, as illustrated in (15) below:

(15) Dominique Strauss-Kam, the French head of the International Monetary Fund, escaped dismissal for a one-night stand with a subordinate today, but was denounced by board members for a "serious error of judgment”. (Report 1)

5.4 Textual and Interpersonal Themes

The analysis of the textual Themes revealed that these are much more frequently used in commentaries (20%) than in news reports (6%), and that these differences are statistically-significant (p<0.05). This means that writers of commentaries rely much more on textual Themes as textual signals to organize the logico-semantic relations in texts, probably due to the complexity of the ideas and arguments that are presented to readers. By contrast, information in news reports can be presented in a more straightforward way with the help of other textual devices such as paragraphing.

Interpersonal Themes are almost non-existent in both groups of texts. In commentaries very few items were found (‘of course’, ‘true’, ‘admittedly’). This can be due to the fact that commentaries express interpersonal meaning through other linguistic devices, and do not rely so much on thematisation for this purpose.

6 Discussion and concluding remarks

The comparative analysis of the thematic selection and realization features of the sample of news reports and commentaries used in this study has yielded a number of interesting findings which deserve the attention of the analyst in search for linguistic signals of a text’s genre.

As the analysis has shown, the writers of news reports and commentaries choose different thematic features to structure their messages, preferring certain types over others. Thus, we have seen that in news reports the preferred type of experiential role selected as Thematic Head is the *Sayer* in a verbal process. This preference, in our view, is the result of the newsmaker’s decision to attribute information to outside sources to give an impression of factuality and objectivity. This choice contrasts with preferred use of the *Carrier* in a Relational process in commentaries, a textual strategy to present the writer’s views as unattributed evaluations, as opinions based on facts.

The analysis of the types of Nouns selected as Thematic Heads also reveals interesting differences, which can be interpreted as signals of the different types of subject matter which characterises these two newspaper genres. News reports are usually more about concrete topics and this is reflected in the higher frequency of concrete nouns, referring to individuals, groups of people or institutions. By contrast, commentaries are more about abstract issues, and this is reflected in the higher frequency of abstract nouns. Many of these abstract nouns are nominalizations, a grammatical device that allows the writer to detach himself/herself from the situation, thus giving the impression of impartiality when expressing his/her own views.

As to the internal structure of the Nominal Groups as Thematic Heads, it was observed that they tend to be shorter in news reports, and often clarified through the use of Appositions which clearly identify the referent. This is due to the need to provide a high degree of truth and clarity. In commentaries, by contrast, they tend to be longer and more complex, frequently followed by Qualifiers. The impression is one of academic, formal discourse, of a more elaborated style than that of news reports, which rely more on verbs than on nominals for conveying meaning.

The low frequency of interpersonal Themes indicates that writers of these two written newspaper genres prefer to use other linguistic means –not Theme- for expressing interpersonal meanings. A cursory analysis reveals the use of modality, involvement strategies such as the use of 'we' in commentaries to involve the general public, conditional sentences, and evaluative lexis.

Interestingly, the different frequencies observed in the use of textual Themes is a clear signal of the different textual structures which characterize these two genres. In news reports textual Themes are not frequent since the textual organization relies on paragraphing. Each paragraph reports a finding or a newsmaker's comment. By contrast, textual Themes are a fundamental tool for writers of commentaries. They rely on textual Themes systematically to scaffold the text's argumentative structure, and to signal logico-semantic relations between complex ideas.

In conclusion, the micro-thematic analysis carried out in this study has revealeinteresting differences between news reports, on the one hand, and commentaries, on the other, which can be interpreted as linguistic signals of the different genres to which these texts belong. Further work at a macro-structural level (e.g. thematic progression patterns, distribution of themes in text stages) is currently under way, and will, hopefully, complement the results of this initial study.

References

ALONSO, I. (coord.) (1997). Different Approaches to Newspaper Opinion Discourse.

Special Issue of Revista Electrónica de Lingüística Aplicada. Asociación Española de Lingüística Aplicada (AESLA).

BELL, A. (1991). *The language of News Media*. Oxford UK & Cambridge MA: Blackwell.

BHATIA, V.K. (2002). A Generic View of Academic Discourse. In J. Flowerdew (ed.) *Academic Discourse*. Harlow, London and New York: Longman. 21-39.

DAVIES, F. (1989). *Reading between the lines*. ELR Seminar Paper. University ofBirmingham.

DAFOUZ, E. (2008). The pragmatic role of textual and interpersonal metadiscourse markers in the construction and attainment of persuasion: A cross-linguistic study of newspaper discourse. *Journal of Pragmatics* 40 (2008): 95-113.

EILER, M. (1986). Thematic distribution as a heuristic for written discourse function. In B.

Couture (ed.). *Functional Approaches to Writing, Research Perspectives*. Norwood, New Jersey: Ablex.

FAIRCLOUGH, N. (1995). *Media Discourse*. London and New York: Edward Arnold.

FRANCIS, G. (1989). Thematic Selection and Distribution in Written Discourse. *Word,* 40,

No. 1-2, 201-222.

FRANCIS, G. (1990). Theme in the Daily Press. *Occasional Papers in Systemic Linguistics,* 4, 51-87.

Fries, P. H. and G. Francis (1992). Exploring Theme: Problems for Research. *Occasional*

Papers in Systemic Linguistics Volume 6, 45-59.

Hawes, T. & Thomas, S. (1996). *Rhetorical uses of theme in newspaper editorials*. In World Englishes 15(2), 159-170.

Lavid, J. (2000). Contextual constraints on thematisation in discourse: an empirical study. In P. Bonzon, M. Cavalcanti and R. Nossum (eds.). *Formal aspects of context*. Dordrecht: Kluwer Academic Publishers, 37-47.

Lavid, J., Arús, J. and L. Moratón (2009). Comparison and translation: towards a combined methodology for contrastive corpus Studies. *International Journal of English Studies*. *Special issue on Recent and Applied Corpus Studies*. Pascual Cantos & Aquilino Sánchez (eds). 159-173.

Lavid, J., Arús, J, and JR Zamorano (2010). *Systemic-functional grammar of Spanish: a contrastive account with English*. London: Continuum.

Nwogu, K. And T. Bloor (1991). Thematic Progression in Professional and Popular Medical

Texts. In E. Ventola (ed.) *Functional and Systemic Linguistics: Approaches and Uses.*

Mouton de Gruyter, 369-384.

Scollon, R. (1998). *Mediated Discourse as Social Interaction: A Study of News Discourse*. London and New York: Longman.

Tirkkonen-Condit, S. (1996). Explicitness vs. explicitness of argumentation: and intercultural comparison. *Multilingua* 15 (3), 257–273.

van Dijk, T.A. (1988). *News as Discourse*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, W. (2008). Intertextual aspects of Chinese newspaper commentaries on the events of 9/11. *Discourse Studies* 10 (3), 361-381.

White, P.R.R. (1997). Death, disruption and the moral order: the narrative impulse in mass ‘hard news’ reporting. In Christie, F. & Martin, J.R. (eds.) *Genre and Institutions: Social Processes in the Workplace and School*. London and Washington: Cassell, 101-133.

On the signalling of multi-level discourse structures

Lydia-Mai Ho-Dac(1), Cécile Fabre, Marie-Paule Péry-Woodley, Josette Rebeyrolle(2)

(1) VALIBEL, Université Catholique de Louvain (Louvain-la-Neuve)
lydia.ho-dac@uclouvain.be
(2) CLLE-ERSS UMR5263, Université de Toulouse - UTM
{fabre,pery,rebeyrolle}@univ-tlse2.fr

Abstract In this paper, we present a top-down approach to discourse annotation which considers the text as a whole as a functional and semantic unit and focuses on the identification of high-level textual patterns. Two discourse objects are chosen to bootstrap this global approach to discourse: *enumerative structures* and *topical chains*, envisaged as "neutral" structures covering a wide range of discourse organisation phenomena and realised via diverse textual patterns. We present here the NLP-assisted annotation method that has been carried out, involving automatic tagging and pre-marking and the use of a purpose-built annotation interface.

Keywords: Discourse organisation, Corpus annotation, textual patterns

1 Introduction

Research focusing on signals of text organisation is of necessity rooted, whether implicitly or explicitly, in a particular view of discourse structure and of discourse segments. A most intuitively attractive view – and one which has a vast potential for application in a diversity of fields – is that texts are organised in terms of major topics or themes. This is the view which underlies the various approaches to automatic text segmentation, which exploit automatically detectable breaks in lexical cohesion to identify topical shifts (Hearst 1997). Yet the notion of discourse topic is far from being an area of theoretical consensus. What relation do discourse topics entertain with sentential or propositional topics? Should they be thought of in terms of a summarizing proposition (cf. van Dijk 1977)? Or rather in terms of major participants (Givón 1983)? This approach in terms of topic continuity is connected to the definition of text segments on the basis of topical chains (Cornish 1998; 1999). But such conceptions of text organisation around topical referents are a long way from "rhetorical" conceptions, where segments are arguments entering into discourse relations – e.g. result or contrast –, and where connectives are the archetypical “signals”. A more classically rhetorical approach views texts as composed of segments belonging to different genres: typically a work of fiction will alternate descriptive and narrative sections, an expository text descriptive and argumentative ones. Specific signalling strategies can be associated with these textual strategies: temporal and spatial adverbials have attracted much attention (Virtanen 1992; Piérard & Bestgen 2008). Yet another take on the question of text organisation focuses on “document structure” (Power et al. 2003) or text architecture (Luc et al. 1999); for these authors, segments include chapters, sections and sub-sections, paragraphs; signals include layout features, headings, and also possibly so-called “metadiscursive” elements.

The purpose of this rapid and superficial survey is to illustrate the complexity of text organisation: different processes are going on concurrently, calling upon a variety of signalling resources, many of which may be shared by several processes. Titles are a case in point: as identifiable boundaries between chapters or sections they are clearly involved in signalling document structure, yet in most cases (except titles such as “Chapter 2”), they also introduce textual contents (thematic or referential, see Rebeyrolle et al. 2009). An important aspect of this complexity is that, contrary to the image called to mind by the term “text-tiling” (Hearst 1997), text structure cannot be thought of as flat. All the approaches above take into account the possibility of nesting: topics within topics, relations within relations, descriptive sections embedded within narratives, sub-sections within sections. This recursive property implies another: the various structuring modes or principles apply at different levels of granularity, involving spans of text of widely different sizes.

The possible “isomorphism” between segments defined by these different approaches is an open question: the correspondence between paragraphs (elements of document structure) and thematic segments for example, which is often taken for granted in writing manuals, is far from established. Similarly, techniques calling upon connectives or temporal adverbials to reinforce a text-tiling approach to text segmentation make the assumption that there is a correspondence between rhetorical and thematic segments, a correspondence which cannot be taken for granted.

To sum up this complex picture, it is possible to identify three basic properties of text:

- text organisation is *multi-dimensional*, and text is the result of a “struggle between (…) different forces” (Enkvist 1985). Systemic Functional Linguistics usefully conceptualises this multi-dimensionality in terms of three metafunctions: ideational, textual, interpersonal;
- most text organisation principles function recursively, leading to nested structures;
- most structures are *multi-level*, in the sense that they can be found at different levels of granularity.

Finally, the signalling resources themselves exhibit a twofold complexity (besides being polysemous) : a) many are *multi-functional*, i.e. they work on several planes at once (cf. headings above); b) few are discrete linguistic elements, and it is generally more useful to think of signals as *configurations* or *bundles* of cues (e.g. temporal adverbials in paragraph initial position, cf. Ho-Dac & Péry-Woodley 2009).

We propose in this paper an approach designed to advance the description of multi-level discourse structures and their signalling while taking into account these various areas of complexity.

2 Two multi-level discourse structures

In order to retain a manageable research programme without over-simplifying the issues, we restrict our investigations to two basic discourse structures: enumerative structures and topical chains. These two structures, which are very frequent in expositive texts, have been selected because they exemplify in interesting ways the properties referred to above; in particular, they are both multi-level.

2.1 Enumerative structures (ESs)

Enumerative structures (ESs) are a prime tool to implement the linearity constraint, setting out a range of discourse elements in the linear format of written text. They have the capacity to realise (dis)continuity strategies into a textually coherent segment, i.e. a unit with boundaries (discontinuity with surrounding text) and an internal organisation (internal continuity). This textually coherent segment prototypically includes three kinds of sub-segments: a trigger (i.e. an introductory segment) announcing the enumeration, an enumeration corresponding to a list of at least two co-items which may enter into various discourse relations; a closure which ends the enumeration.

Example 1 is a complex ES composed of four co-items (3.1 to 3.4), with what looks like a fifth item functioning in fact as a closure (3.5 /Rapprochements/ ‘‘Bridges’’), as shown by the encapsulating NP /Les approches explicitées ci-dessus/ “The approaches described above”. The whole ES stretches over several paragraphs, taking up an entire subsection, and contains several embedded ESs. The top-level co-items are clearly linked to document structure, as each corresponds to a subdivision of section 3 (/3. Fondements sociaux du concept en occident/ “3. Social foundations of the concept in the West”).

While the enumeration is a necessary component of ESs, triggers and closures are optional. Our annotation scheme also includes the enumeraTheme (in bold in example 1): the optional explicit expression of the similarity criterion which underlies an enumerative structure, the co-enumerability criterion[1]. The notion of enumeraTheme is in partial overlap with the notion of hyperTheme and hyperNew developed by Martin and Rose:

[1] “The textual act [of Enumerating] consists in transposing textually the co-enumerability of the listed entities into the co-enumerability of the linguistic segments describing them, which thereby become the entities constituting the enumeration (the items). The identity of status of the items in the enumeration expresses the identity of status of the listed entities in the world.” (Luc *et al.*, 2000: 25, our translation).

> While hyperThemes predict what each phase of discourse will be about, new information accumulates in each clause as the phase unfolds. In written texts in particular, this accumulation of new information is often distilled in a final sentence that thus functions as a hyperNew to the phase. HyperThemes tell us where we're going in a phase; hyperNews tell us where we've been. (Martin and Rose, 2003: 182)

The enumeraTheme is a broader notion, however: a) ESs are of interest to us as basic structures which can be used to organise different dimensions of discourse presentation. The enumeration principle can be turned into topical, temporal or rhetorical structures by organising discourse hierarchically in terms of sub-topics, properties, events, processes, arguments, etc.; we therefore expect to be able to identify different categories of enumeraThemes corresponding to these different functions; b) enumeraThemes are largely independent of their textual realisations, including their textual position: preceding the enumeration (Theme) or following it (New).

ESs, their components and their signalling have been the object of a number of studies (Turco and Coltier, 1988; Adam and Revaz, 1989; Luc et al. 2000; Péry-Woodley, 1998 and 2001; Maurel *et al.*, 2003; Jackiewicz 2005; Porhiel 2007 *inter alia*), which provide the foundations for our current project. These studies tend however to focus on prototypical signals of ESs: discourse markers, including, in some cases, visual properties. Our own focus on the text organisational properties of the structure leads us to wish to take into account a broader range of signals. This diversity of functions and of realisations is what makes enumerating a fundamental discourse organisational device. To address this two-fold diversity, an empirical approach seems necessary: starting from attested structures, and calling upon corpus-linguistics methods and machine learning techniques to characterise them both in terms of function and signalling. The starting point of such a methodology is a human annotation campaign, involving in our case a certain amount of computer-based pre-processing and assistance (sections III.1 and 2). Only then will it be possible to set up procedures to identify signals and characterise different functions (section III.3).

		embedded ESs
3. Fondements sociaux du concept en occident	**TRIGGER**	
3.1. Les principes moraux	**ITEM 1**	
Deux **principes** liés au sujet font consensus dans les démocraties occidentales :		**TRIGGER**
1. la réprobation de la contrainte sur autrui (ou « atteintes à la personne humaine ») : […]		**ITEM 1**
2. le respect de la volonté individuelle : […]		**ITEM 2**
Un autre principe est lui assez largement partagé : la protection de l'enfant comme volonté positive affirmée : […]		**ITEM 3**
3.2. Le point de vue du droit	**ITEM 2**	
Dans l'étendue des choses qu'il traite en matière de droit des personnes, et pour éclairer le sujet, le droit peut s'analyser en deux **notions**		**TRIGGER**

complémentaires :		
1. ce qui est de l'ordre du sujet : […]		**ITEM 1**
2. ce qui est de l'ordre de l'objet : […]		**ITEM 2**
Cette analyse dichotomique appelle trois **remarques** :		**CLOSURE + TRIGGER**
Elle pose la question de la frontière entre sujet et objet. Elle se résout par […]		**ITEM 1**
En matière de relations sexuelles, elle fait de l'adulte un sujet, […]		**ITEM 2**
Le droit pénal condamne des actes, qu'ils soient délictueux ou criminels : […]		**ITEM 3**
3.3. Le point de vue médical	**ITEM 3**	
La médecine contribue au débat en répondant à la question de la maturité biologique par rapport à la sexualité. […]		
3.4. Le point de vue psychologique	**ITEM 4**	
Au-delà de la simple maturité du corps envisagée par la médecine, se pose la question de la maturité psychologique de l'individu. C'est une notion assez vague, où l'on peut distinguer deux **aspects** :		**TRIGGER**
3. la maturité sociale, c'est-à-dire la capacité de […]		**ITEM 1**
4. la maturité sexuelle, ou en d'autres termes la capacité […]		**ITEM 2**
Ce qu'on peut en tout cas affirmer sur les deux **alinéas** précédents, c'est qu'ils sont très dépendants de l'éducation et des circonstances de vie de chacun. […]		**CLOSURE**
3.5. Rapprochements	**CLOSURE**	
Les **approches** explicitées ci-dessus forment l'essentiel des principes qui justifient la manière dont nos sociétés perçoivent la pédophilie et sa pratique. […]		

Example 1: A complex enumerative structure

2.2 Topical chains (TCs)

Topical chains are the second type of object in our study. Like ESs, TCs can organise text at different levels of granularity: from local units, consisting of a few sentences, to whole texts (see example 2 below, where the whole of section 3 (*Personality*) is organized around one topical referent, Albert Einstein, which is also the starting point of topical chains in other sub-sections). A topical chain is a text segment by virtue of the links established from unit to unit in a string by linguistic expressions referring to an entity initially established as topic. It is therefore a specific form of cohesive chain. TCs are likely to combine various types of referential expressions, such as proper names, definite NPs, demonstrative NPs, etc. and anaphoric reference via pronouns. Because of the strong association between sentence topic and grammatical subject, we focus on referential and anaphoric expressions in subject function (in bold in example 2, where the referent is repeted via NPs reiteration, possessive NPs, anaphoric pronouns).

Discourse research has taken an active interest in cohesive chains, and to a lesser extent in the more restrictive topical chains. Many studies have focussed on the description of co-referential expressions and anaphora phenomena (Ariel 2004; Charolles 2002; Cornish 1998; Schnedecker 2005, inter alia), others on corpus annotation and the automatic detection of co-referential chaining. Quite a few difficulties subsist, however, even with human identification of thematic dis/continuities (cf. Biber et al. 2006, chapter 20). We see in TCs a distinct form of textual organisation based on different strategies and resulting in different textual patterns from enumerative structures. Our model concerning TCs is however less developed at this stage than for ESs, and their annotation approached in an exploratory mode.

3. Personnalité

3.1. Einstein et la politique

Les positions politiques prises par Einstein sont marquées par ses opinions pacifistes, qu'il relativise parfois, […]. En 1913, **il** est cosignataire d'une pétition pour la paix que trois autres savants allemands acceptent de signer. […]

3.2. Vie sociale

Einstein a rencontré un grand nombre de personnalités majeures de son époque, […].

3.3. Einstein et la religion

Einstein écrit plusieurs textes traitant des relations entre science et religion. Dans son article paru en 1930, **Einstein** distingue trois formes de religion […]

3.4. Einstein et la philosophie

La philosophie n'est pas l'un de ses domaines de prédilection, mais **Albert Einstein** marque son intérêt pour […].

3.5. Einstein et l'astrologie

Contrairement à la citation qui lui est attachée par de nombreuses publications, en particulier celle de l'astrologue Élizabeth Teissier, **Einstein** ne croyait pas en l'astrologie. […].

3.6. Einstein et le végétarisme

Albert Einstein soutient la cause végétarienne. **Il** considère le végétarisme comme un idéal sans pourtant […]. **Ses arguments** se basent principalement sur des raisons de santé, […].

Example 2: A Topical Chain extending over a whole section

This investigation of enumerative structures and topical chains is being developed within the ANNODIS project[2]. The general objective of the project is to construct an annotated corpus of French-language written texts, to be made available via an interface for the study of discourse organisation. In this paper we focus on the design and setting up of the human annotation campaign for our two structures, in relation to the planned exploitation of annotations. The signalling of text organisation is at the heart of both these steps, and the approach for both relies quite heavily on NLP techniques: in the annotation step, coders are guided by

[2] A multidisciplinary project funded for three years by the French National Research Agency (ANR). URL: http://w3.erss.univ-tlse2.fr/annodis/.

automatically identified candidate cues which are highlighted in the text; in the exploitation step, annotations are analysed in terms of statistically significant correlations between cues, and between cue configurations and structures.

3 The signalling of multi-level discourse structures

In order to study these complex signalling mechanisms, we have developed an approach inspired from Biber’s methodology for an emergent text-typology (Biber 1988). In this case, discourse markers will emerge from the analysis of regularly co-occurring text features. In our approach, a discourse marker should not necessarily be thought of as a discrete linguistic element (typically, a connective or an adverbial), but rather as a configuration of cues – lexico-syntactic, positional, etc. This hypothesis arises from previous work, in particular Ho-Dac and Péry-Woodley (2009), where a corpus-based methodology is used to describe the discourse function of time adverbials in initial position. According to this study, a sentence-initial time adverbial does not in itself have a text structuring capacity (as often argued): its capacity to indicate the beginning of a temporally homogeneous discourse segment is linked to specific contexts, such as the start of a new subsection or paragraph, and only when the adverbial is part of a series. Such configurations are strongly text-type dependent.

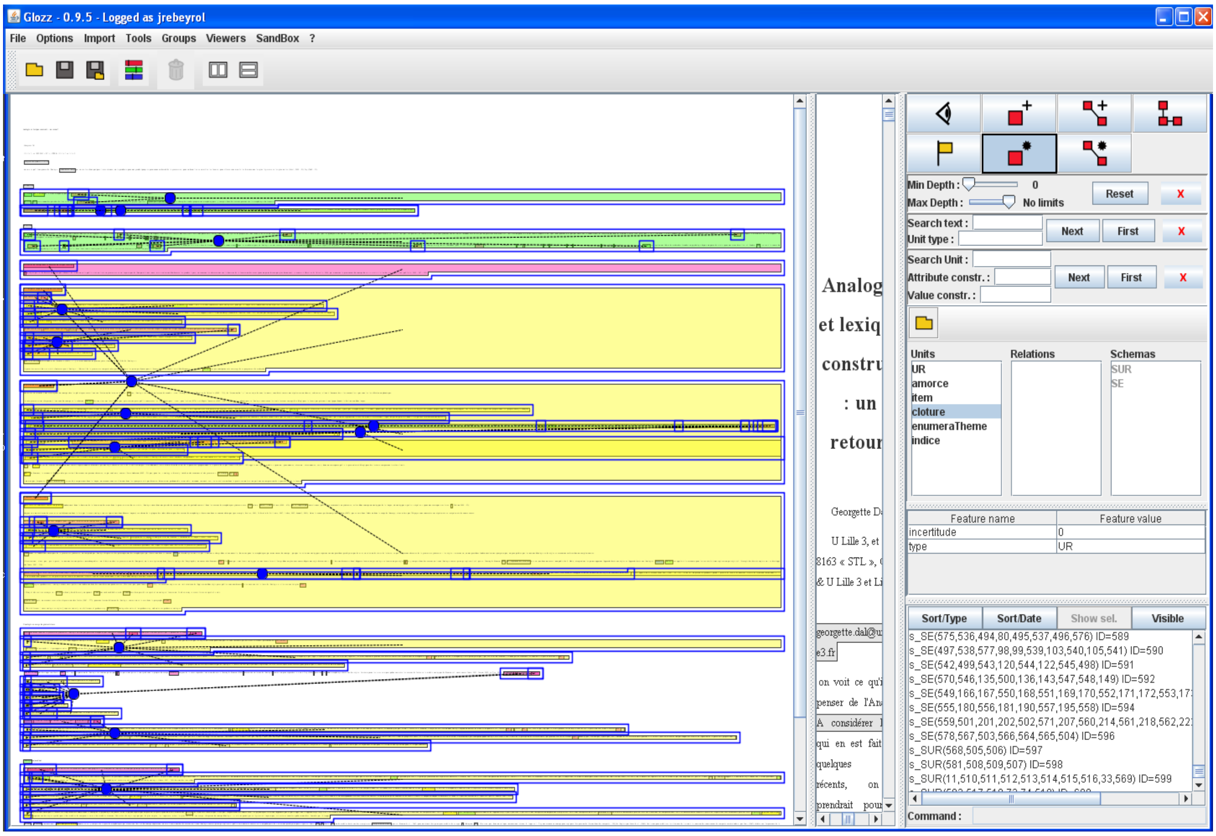
When seeking to identify markers signalling multi-level discourse structures, the features we need to consider will include, along with specific lexico-syntactic elements, text position, patterns of repetition, relationships between words and text structure, text types, etc. This approach calls for a carefully designed experimental framework, which we describe in the next section.

3.1 Experimental framework and Annotation campaign

Texts selected for inclusion in the corpus combine three sub-types of long expository texts: encyclopaedia articles (from Wikipedia), scientific papers (from proceedings of a linguistics conference) and reports in the field of international relations (from the website www.ifri.org). These texts are encoded in XML, following TEI-P5 encoding procedures, and premarked with the features described in the following subsection. Their format is adapted to the GLOZZ annotation interface (Widlöcher & Mathet 2009). The premarking procedure uses the outputs of automatic POS tagging (Treetagger), syntactic dependency analysis (Syntex – Bourigault 2007), and layout analysis (textual positions are identified, such as first or last sentence of a paragraph, sentence following a heading, etc.), and applies local grammars to identify the features.

An annotation guide is available where the different components of an ES and a TC are defined, illustrated and associated with a list of features and with a number of basic tests to help detect them. Coders are guided in their task by a wide range of features highlighted using a colour code, which allows them to spot classes of features.

On the left of the interface screen, a global view of the text appears in a “ribbon” where features are highlighted. The central window provides a readable view of the text where structures can be annotated. The coder uses the ribbon to scan the text. Once a dense zone of premarked features is detected, the coder can zoom on this zone in the central window and check for the presence of a structure. This structure is then annotated by delimiting and characterising the components. Colour-marking of cues and structures allows a different way of accessing text, starting from a global view which encourages a scanning/skimming approach and in due course zooming down to a more local view. Through this dual way of accessing text, the interface guides the annotator towards a top-down approach for scanning and towards a bottom-up approach for annotating, focusing alternatively on global structures and on more local cues. In figure 1 below the 'ribbon' part of the screen has been extended (to the detriment of the middle 'normal text' window) in order to show the visualisation of text structures which is made possible by the GLOZZ interface.



Feature name	Feature value
incertitude	0
type	UR

```
s_SE(575,536,494,80,495,537,496,576) ID=589
s_SE(497,538,577,98,99,539,103,540,105,541) ID=590
s_SE(542,499,543,120,544,122,545,498) ID=591
s_SE(570,546,135,500,136,143,547,548,149) ID=592
s_SE(549,166,167,550,168,551,169,170,552,171,172,553,17
s_SE(555,180,556,181,190,557,195,558) ID=594
s_SE(559,501,201,202,502,571,207,560,214,561,218,562,22
s_SE(578,567,503,566,564,565,504) ID=596
s_SUR(568,505,506) ID=597
s_SUR(581,508,509,507) ID=598
s_SUR(11,510,511,512,513,514,515,516,33,569) ID=599
```

Figure 1: colour marking in the Glozz annotation interface

3.2 Relevant features for premarking

As described above, our automatic premarking is carried out by way of local grammars applied to POS- tagged and syntactically analysed text. These local grammars detect occurrences of text features which have been selected on the basis of previous studies, in particular by Ho-Dac (2007). In our approach to features, we take into account, in order of importance, textual features (typo-dispositional, punctuational, etc.), syntactic features (subject position, type of phrases (NP and PP), etc.) and lexical features (connectives, time nouns, space nouns, adverbs, prospective elements, etc.). Typo-dispositional features taken into account include section headings and paragraph breaks. By positional features, we mean initial detached position, subject position and postverbal position (such as circumstantial adverbials). Recent studies have shown initial position to be highly strategic as it is endowed with an essential function in text processing, the orientation function (cf. Chafe 1994).

Because ESs and TCs (topical chains) appear at very different granularity levels, their signalling involves features which are very diverse and which may be selectively linked to granularity level. The features we are able to automatically premark on a text are not necessarily all relevant for the annotation of ESs and TCs and may disturb the annotation task. Among the three different types of features mentioned above, a selection of the ones which are directly relevant for the annotation of the two structures is presented in table 1 below.

Features	**Description**
Lexico-syntactic features	spatial adverbs or spatial PP temporal adverbs or temporal PP notional adverbs or notional PP connectives
Prospective elements	NP including a lexical item likely to signal a prospective element (*les suivants…* [The following ...])
Encapsulations	demonstrative NP including a lexical item and a numeral likely to signal an encapsulation (*ces trois scenarios…* [These three scenarios ...])
Adverbial textual organisers	*premièrement, parallèlement, enfin,* etc. [First, At the same time, finally]
Formatted lists	Textual structures associated with visual properties.
Layout features	headings paragraph breaks bullets indentation
Punctuational patterns	The last word before ':' punctuations marks coordinators in sentences including ': . , ; et ou'
Positional features	initial detached position subject position postverbal position

Table 1: Premarked features used to assist the annotation procedure

3.3 From features to cue configurations

Once the automatic premarking is completed, we can move on to the annotation phase. Within the annotation process, the premarked features are seen as *candidate cues* meant to help the annotators adopt a more global view of the text, and to facilitate the identification of sporadic discourse structures. These candidate cues will only become fully-fledged cues when they have been validated by the coders; they will then be associated with the structural element in which they occur (trigger, item, closure, topical chain). The signalling of a structure or a structural element is generally not achieved by a single cue but by several cues interacting and forming a *cue configuration*.

A typical example of *cue configuration* in an ES is an NP containing a time noun *or* a textual organiser (*premier, second*, etc.) *or* a PP introduced by a temporal preposition (such as *depuis, durant, dès*, etc.) in initial detached position after a paragraph break and followed by another noun or prepositional phrase containing a similar lexical item.

Our final objective within the ANNODIS project is to exploit the annotated corpus with text mining procedures and make these complex configurations emerge. More than 800 ES are already available for analysis on the basis on the first 65 annotated texts.

These data will allow us to extract combinations of cues associated with these structures, and in particular to examine the variety of signalling devices depending on the granularity of the structures, differentiating between structures which occur within the limits of the paragraph and structures that stretch over large spans of texts. These cue configurations are the discourse markers which signal our structures.

Alongside sophisticated text mining techniques, the interrogation interface will also allow linguistic research with more traditional tools such as concordancers, and will make it possible to enter additional annotations, whether manually or through the projection of local grammars.

Acknowledgements

We would like to thank Ludovic Tanguy for the data analysis.

References

Ariel, M. (2004). Accessibility Marking: Discourse Functions, Discourse Profiles, and Processing Cues. Discourse Processes, 37(2), 91-116.

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Bourigault, D. (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Mémoire d'HDR en Sciences du Langage, CLLE-ERSS, Toulouse, France.

Chafe, W. (1994). *Discourse, consciousness and time.* Chicago & London: Univ. of Chicago Press.

Charolles, M. (2002). La référence et les expressions référentielles en français. Paris: Ophrys.

Cornish, F. (1998). Les "chaînes topicales" : leur rôle dans la gestion et la structuration du discours. Cahiers de Grammaire(23), 19-40.

CORNISH, F. (1998). Les chaînes topicales : leur rôle dans la gestion et la structuration du discours. *Cahiers de Grammaire* 23, 19-40.

CORNISH, F. (1999). *Anaphora, Discourse and Understanding. Evidence from English and French.* Oxford: Clarendon Press.

ENKVIST, N.E. (1985). A parametic view of word order. In E. Sozer (ed.) *Text Connexity Text Coherence: Aspects Methods Results.* Helmut Bushe: Hamburg, pp.320-336.

GIVÓN, T. (1983). Topic continuity in discourse: an introduction. In T. Givón (ed.) *Topic continuity in discourse: a quantitative crosslanguage study.* John Benjamins: Amsterdam/Philadelphia, pp.1-42.

HEARST, M.A. (1997). TextTiling : Segmenting Text into Multi-parapgrah Subtopic Passages. *Computational Linguistics*, 23 (1), 33-64.

HO-DAC, L.-M. & PÉRY-WOODLEY, M.-P. (2009). A data-driven study of temporal adverbials as discourse segmentation markers. *Discours*, 4, special issue 'Linearization and Segmentation in Discourse'.

HO-DAC, L.-M. (2007). *Exploration en corpus de la position initiale dans l'organisation du discours*. Thèse de doctorat en sciences du langage. Université de Toulouse 2.

LUC, C., MOJAHID, M., VIRBEL, J., GARCIA-DEBANC, C., & PÉRY-WOODLEY, M.-P. (1999). A linguistic approach to some parameters of layout: A study of enumerations, *AAAI 1999 Fall Symposia "Using Layout for the Generation, Understanding or Retrieval of Documents"*, North Falmouth, Massachussetts (1999), p. 20-29.

PIÉRARD, S. & BESTGEN, Y. (2008). Use of temporal adverbials as segmentation discourse markers by second language learners. *Archives de Psychologie, 73*, 209-230.

POWER, R., SCOTT, D. & BOUAYAD-AGHA, N. (2003). Document structure. *Computational Linguistics*, 29 (2), 211-260.

REBEYROLLE, J., JACQUES, M.-P. & PÉRY-WOODLEY, M.-P. (2009). Titres et intertires dans l'organisation du discours. *Journal of French Language Studies*, 19 (2), 269-290.

SALMON-ALT, S. (2002). Le projet Ananas : Annotation Anaphorique pour l'Analyse Sémantique de Corpus. Paper presented at the Workshop sur les Chaînes de référence et résolveurs d'anaphores, TALN 2002, Nancy, France.

SCHNEDECKER, C. (2005). Les chaînes de référence dans les portraits journalistiques : éléments de description. Travaux de Linguistique, 51, 85-133.

VAN DIJK, T.A. (1977). *Text and Context. Explorations in the Semantics and Pragmatics of Discourse*. London: Longman.

VIRBEL, J. (1985). Langage et métalangage dans le texte du point de vue de l'édition en informatique textuelle. *Cahiers de Grammaire,* 10, 5-75.

VIRTANEN, T. (1992). Given and new information in adverbials: Clause initial adverbials of time and place. Journal of Pragmatics 17 (2), 99-117.

WIDLÖCHER, A. & MATHET, Y. (2009) La palte-forme GLOZZ: environnement d'annotation et d'exploration de corpus. TALN 2009, Senlis.

Lexico-grammatical Discourse Features of Interdisciplinary and Interprofessional Co-operation

David Oakey (1), Peter Mathias (2)

(1) Iowa State University, USA
djoakey@iastate.edu
(2) Bridge Research and Development, UK
petercmathias@btinternet.com

Abstract

This paper presents a comparative corpus-driven study of epistemic signaling devices in interdisciplinary research articles. Much has been written on the language of research articles in separate disciplines, but few descriptions currently exist of the language used in collaborative research between practitioners from different disciplines. Furthermore, while it has been pointed out that potential research collaborators from different fields need to become familiar with the language used in each others' disciplines (Committee on Science, 2004), there is little guidance on which linguistic features should take priority in any familiarization program. This paper explores two corpus-driven methodological approaches to identifying textual features which signal the epistemological basis of interdisciplinary and contributory disciplines. Data are taken from journals in the interdisciplinary field of Interprofessional Care and its contributory disciplines, Medicine and Social Work. The first approach uses frequently occurring lexical bundles (Biber et al, 1999), while the other focuses on collocations of salient grammatical words (Groom 2007). We show that both approaches reveal similarities and differences between the epistemologies of the interdisciplinary field and its contributing disciplines, and discuss potential applications of these findings.

Keywords:

academic discourse; interdisciplinary collaboration; lexical bundles; salient grammatical words

1. Introduction: Epistemology and interdisciplinary cooperation

The idea that the search for new knowledge can progress more swiftly through collaboration between scholars in different disciplines is not new. For example, the literature on interdisciplinarity comprises several decades of sociological studies of cooperation (or the lack of it) between academic "tribes" (Campbell 1969/2005; Becher, 1989). Much effort has also been devoted to facilitating collaboration by researchers across disciplines, primarily focusing on the identification and removal of institutional impediments to closer collaboration, such as the difficulty of obtaining funding, and physical distance between collaborators on campus or further afield (Derry and Schunn, 2005). Surveys of collaborating scholars, such as that by the US National Academy of Sciences (Committee on Science, 2004), have shown that "without special effort by researchers to learn the languages and cultures of participants in different traditions, the potential interdisciplinary research might not be realized and might have no lasting effect" (ibid., p21). Furthermore, although the committee asserts that "in their written and oral communications, researchers and faculty members can facilitate IDR by using language that those in other disciplines are able to understand" (ibid., pp81-82), there is a dearth of studies of the specific linguistic aspects of interdisciplinary work which might facilitate such interdisciplinary research efforts.

This paper consequently reports an exploratory study into some of the linguistic aspects of interdisciplinary text. Specifically, the research focused on linguistic features that reveal or signal epistemology, the underlying foundational concepts, principles and practices that are taken as given in a particular discipline. These epistemological features are of particular interest and importance in interdisciplinary research endeavors, which as noted above, require skillful cooperation between the practitioners of a number of disciplines. Readers from one discipline may need a grasp of epistemology of another discipline to understand the possibilities of cooperation fully and to identify points of potential misunderstanding and conflict, although the extent to which participants require a deep understanding of contributory disciplines may depend on the type of cooperation involved[1]. The paper explores whether or not it might be possible to identify features of text which signal epistemology as a step towards developing a new method to support those participating in or leading interdisciplinary initiatives.

2. Signaling epistemology

2.1 Corpus Linguistics, Lexical Bundles and Salient Grammatical Words

In business and academic research, as well as in other areas of linguistic enquiry, there is increasing use of methods developed in corpus linguistics: methods which center on corpus design, construction and interrogation. Corpus linguistics is attractive because its methods potentially allow a combination of the quantitative and qualitative. Analysis can be statistical and/or interpretative and in either case the raw material is in a form that can be shared amongst scholars and other analysts readily and easily making it possible for more than one person to take part in the research or attempt to replicate the results.

[1] For reasons of space we will not draw distinctions here between *multi-*, *inter-*, *cross-* or *trans*disciplinary research. For a full discussion see Mathias (2009).

Lexico-grammatical Discourse Features of Interdisciplinary and Interprofessional Co-operation

The investigative methods described in the paper draw from those used by Biber et al. (2004), Groom (2007) and Oakey (2002; 2009) who use corpus-driven approaches to identify structuring devices in academic discourse. Biber et al. (1999) used *lexical bundles*, frequently occurring fixed word strings, to reveal linguistic differences between the register of academic prose and the register of conversation. Their principal finding was that the structures of lexical bundles differ across registers: “most lexical bundles in conversation are building blocks for verbal and clausal structural units, while most lexical bundles in academic prose are building blocks for extended noun phrases or prepositional phrases” (ibid., p992).

Of more relevance to the present study was the functional framework developed by Biber et al. (2004) which attempted to categorize the textual functions of lexical bundles. This had three core categories of function which relate to Halliday’s (1994) metafunctional categories of register, namely Stance bundles (corresponding to Halliday's Interpersonal metafunction) which "express attitudes or assessments of certainty that frame some other proposition" (Biber et al. 2004, p384) such as *the fact that the* and *it is necessary to*; Referential bundles (corresponding to Halliday's Ideational metafunction) which "make direct reference to physical or abstract entities, or to the textual context itself" (ibid.), such as *in the absence of* and *the extent to which*. Finally, Discourse Organizers (Halliday's Textual metafunction) "reflect relationships between prior and coming discourse" (ibid.), such as *on the other hand* . The latter two categories, useful for comparing between registers, offer a way to compare the discourse of interdisciplinary collaboration with that of the discourse in the disciplines in which the collaborators are based. Referential bundles may reveal aspects of the epistemology of the interdisciplinary field, while Discourse Organizing lexical bundles could uncover differences in signaling text structure.

The second feature used in this investigation is the *semantic sequence* (Groom 2007), which Groom uses to advance his idea that ‘phraseology and epistemology are mutually constitutive; epistemology is manifested in phraseology, and phraseology produces and reproduces epistemology’ (2007, p6). For Groom, disciplinary discourse is ‘a stable yet continually evolving set of meanings, values and practices which produces and is produced by a stable yet continually evolving set of conventional linguistic forms ... [it] is the fusion of epistemology and phraseology’ (ibid., p25).

The advantage of the semantic sequence, like the lexical bundle, is that it is a corpus-*driven* feature which emerges from the data, rather than being found as the result of a search for any particular item. Corpus-*based* studies, on the other hand, typically interrogate the data on the basis of predetermined categories or concepts, usually involving the use of open class, content or lexical words of some kind such as nouns or verbs. Semantic sequences, ‘repeated sequences of meanings which may be realized through a range of different grammatical forms’ (ibid., p83), are instead indentified by searches for salient grammatical words such as prepositions, determiners, or articles. Any sequence that appears twice in a set of 100 concordance lines is a semantic sequence, such as ‘Property + of + Phenomenon (e.g. *The essential values of academic life*) or ‘Process + of + Object’ (e.g. *The building of a new church*). It is suggested here that semantic sequences may reveal or elicit epistemologies, foundational concepts and characteristic ways of building knowledge in different disciplines, moving the analysis beyond "beyond trivial features of content and towards the identification of deeper conceptual linkages between form and meaning" (ibid., p284).

3. Signaling Epistemology in Health and Social Care

3.1 The interdisciplinary corpus

The field of Health and Social Care is an interdisciplinary field tackling issues requiring the co-operation of researchers and professionals from Medicine and Social Work. A sample of this discourse, consisting of articles in electronic format, was collected from each of three journals: 100 articles from a journal in each of the contributing fields, namely the *British Medical Journal* (BMJ) and the *British Journal of Social Work* (BJSW) and then 100 from the interdisciplinary journal to which researchers in these two fields contribute, the *Journal of Interprofessional Care* (JIC). The relative size of the subcorpora is shown in Table 1 below:

Journal	BMJ	BJSW	JIC
Number of articles	100	100	100
Number of tokens	452,497	808, 869	563,055

Table 1: Composition of the Interdisciplinary Corpus

The number of tokens in each subcorpus obviously differs in view of the variation in average text length, but since this is a study of discourse, it is more appropriate to conduct an *isotextual* comparison (Oakey, 2009) in which the subcorpora being compared each contain the same number of communicative acts, and will consequently contain the same number of introduction sections, results sections and so on.

3.2 Lexical bundles

The twenty most frequent 3-word lexical bundles in each subcorpus are shown in Table 2 below. The effect of text length is noticeable: more lexical bundles occur in 100 BJSW articles than in 100 BMJ articles because articles in the former journal are longer than those in the latter journal. The most obvious finding is that many of the lexical words in these strings relate to the subject matter of each discipline, such as *disease* and *patients* in Medicine; *social*, *children*, and *health* in Social Work; and *health* and *social* in Interprofessional Care. In a lexical bundle analysis, then, there are a number of Referential bundles which reflect the concerns and knowledge-making processes of the disciplines. And yet, far more bundles appear to be discipline-neutral, such as *the effects of*, *the use of*, *the need for* and so on. The bundles which most obviously resemble Biber et al.'s Discourse Organizers, such as *as well as*, *in terms of*, and *in order to*, occur frequently in Social Work and Interprofessional Care, but not in Medicine.

BMJ	Freq.	Texts	BJSW	Freq.	Texts	JIC	Freq.	Texts
quality of life	164	17	*of social work*	946	95	*health and social*	399	64
the risk of	124	37	*department of health*	545	57	*and social care*	345	55
the number of	116	42	*in social work*	473	64	*department of health*	311	51
of the study	113	53	*per cent of*	331	72	*as well as*	213	70
in the united	110	42	*social work and*	289	63	*the development of*	185	71
the use of	106	50	*of social workers*	276	100	*in order to*	179	64
body mass index	98	14	*as well as*	273	84	*the need for*	133	59
in patients with	94	30	*social work practice*	261	61	*one of the*	127	64
the effect of	93	45	*in terms of*	259	76	*of health and*	124	53

coronary heart disease	89	16	*social work education*	250	35	*of health care*	120	46
the united states	83	34	*for social work*	243	58	*in terms of*	117	58
the proportion of	82	35	*children and young*	229	29	*the importance of*	110	57
the effects of	75	34	*and young people*	224	26	*a number of*	109	49
the end of	73	27	*the use of*	222	69	*in health care*	107	46
randomised controlled trials	67	17	*a number of*	221	76	*mental health services*	100	18
at follow up	66	13	*in order to*	219	80	*primary health care*	100	30
adjusted for age	66	18	*in relation to*	218	67	*in the community*	100	40
the united kingdom	64	27	*for education and*	198	32	*a range of*	100	43
the quality of	61	21	*health and social*	197	55	*part of the*	99	53
more likely to	61	30	*in the uk*	188	51	*the use of*	94	51

Table 2: 3-word lexical bundles, frequency by journal

These brief results suggest that lexical bundles are an initial indicator of disciplinary epistemologies and text structure, although further detailed work is required to build on these initial findings.

3.3 Salient grammatical words

The salient grammatical words were extracted by comparing the frequency wordlist from the interdisciplinary corpus with a reference wordlist derived from the written component of the British National Corpus (2000). It was possible to identify grammatical words which appear significantly more frequently in the interdisciplinary corpus than in the BNC.

The results of this analysis can briefly be illustrated by four salient grammatical words: *between*, *within*, *of* and *among*. Three 100-line concordance lines were randomly extracted for each word in each journal. Analysis of the four salient grammatical words each offered insights to the epistemology of each field:

Between

Left hand side collocations of *between* featured words of relationship, association, difference, interaction, links communication and collaboration. In terms of frequency, words of relationship and association were most common, followed by words of difference and then words of collaboration, links and interaction in all three corpora. The term *collaboration* appeared almost exclusively in the JIP, whilst the BMJ favored *relationship* and *interaction* and the BJSW *links* and *interaction*. Right hand side collocations of *between* gave an indication of the differences in preoccupation and focus reflected in the three journals: Social Work was concerned with *children, parents, families, local and central government, practitioners* and *professionals*; Medicine with *research, studies, treatment groups, tests,* and *interventions*; Interprofessional Care was more concerned with *teams, organizations, disciplines, stakeholders, professions, interprofessional collaboration , service providers,* and *communities*:

Among

Left hand side collocations of *among* also indicated areas of concern to each discipline: Social Work with *abuse, wellbeing, poverty, stress, trauma*; Medicine with *diseases, causes, survival, mortality, grief, experimentation*; Interprofessional Care with *collaboration, communication, interaction and relationships*. Right hand side analysis of *among* revealed the participants experiencing the phenomena and conditions previously identified in the left hand environments: in the case of Social Work these were *children, older people, families, carers* and *immigrants*; in the case of Medicine these were *infants, children, teenagers, adolescents, women, men and veterans*; in Interprofessional Care *team members, professionals, providers, disciplines, students, the homeless and communities*.

Within

Left hand side collocations of *within* in Social Work illustrate clinical focus, method and process with examples such as: *abuse, adoption, assessment, care, delinquency, dysfunction, fragmentation, distrust, fostering* and *care orders*. Those in Medicine similarly identify focus and process including: *clinical features, conditions* and *interventions, methods of enquiry* and the people and organizations involved. Interprofessional Care shows concern for *role, skills* and *knowledge*, professional activity such as *debate, collaboration* and *communication*, aspects of intervention and the people and organizations involved.

Right hand side words in Social work words describe background and context to thought and action and something of the people involved in practice. Medical concerns are with research outcomes (time), organizations and some clinical features. Interprofessional Care are heavily weighted to organizational context with a lighter emphasis on clinical condition.

Of

Left hand side collocational words of *of* in Social Work show a concern with conceptualization in knowledge building, experiences of clients and the processes of intervention. Collocations in Medicine are concerned with the process of knowledge building, with intervention and with clinical features and processes. Collocations from Interprofessional Care present a more diffuse picture, but it is possible to discern a concern for conceptualization and for the processes of collaboration.

Right hand side words in Social Work deal with experiences and social conditions such as *elder abuse, loss and rejection, being black, learning disabilities, mental illness,* and *solidarity*; they also deal with qualities of intervention in Social Work and Social Care such as *ethics*. Clinical and technical aspects of Medicine feature in right hand collocations along with issues of intervention and knowledge building. Collaborators in Interprofessional Care are concerned primarily with relationships and models of collaboration, cooperation and education.

In general, the collocations occurring immediately to the left and right of each salient grammatical word suggested three dimensions:

- the clinical or academic focus reflecting the concerns and subjects of study, research and practice;
- how knowledge is created within the discipline or profession;
- the organizations and people most closely involved as practitioners or clients.

We suggest that these three dimensions, and specific lexical realizations with each discipline, form a starting point for potential collaborators to familiarize themselves with the language of their collaborating disciplines.

4. Discussion

This investigation shows that it is possible to use lexical bundles and salient grammatical words as probes to discover some of the key features of professional thinking and practice, and to regard them as signaling aspects of epistemology. The epistemological picture that seems to emerge from this brief analysis of lexical bundles and salient grammatical words in the three journals is one in which Medicine is characterized as a relatively hard science, focused largely at the physico–chemical and psychological levels, with significant use of qualitative, experimental methods to advance knowledge in settings mainly dealing with individuals and their family or carer network. Social Work research interests overlap and intersect with those of Medicine at the level of individual and families or carers, but extend more into the psycho-social arena, concerned with the interaction between the individual and society. The interdisciplinary field of Interprofessional Care showed itself to be concerned with the processes of cooperation and collaboration and with the skills, knowledge and competence required to be effective in interprofessional and multi-agency settings.

Naturally, it could be said that these conclusions would have been self-evident to anyone, linguist or not, who took the trouble to analyze the descriptions of role and purpose made available by the relevant professional associations or in the mission statements of the three journals. The counter is of course that the material from the corpus analysis is richer and has the potential, in further work, to reveal more of the underlying concepts, methods and values of the various professions. To collaborating researchers outside linguistics, moreover, the results may indicate that lexical bundles and salient grammatical words offer a new tool to support the organization and management of interprofessional exchange.

References

BECHER, T. (1989). *Academic Tribes and Territories*. Milton Keynes: The Society for Research into Higher Education and Open University Press.

BIBER, D., CONRAD, S., & CORTES, V. (2004). 'If you look at ...': Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, 25(3), 371-405.

BIBER, D., JOHANSSON, S., LEECH, G., CONRAD, S., & FINEGAN, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

British National Corpus World Edition. (2000). Oxford: Oxford University Computing Service.

CAMPBELL, D. T. (1969/2005). Ethnocentrism of Disciplines and the Fish-Scale Model of Omniscience. In S. J. DERRY, C. D. SCHUNN & M. A. GERNSBACHER (Eds.), *Interdisciplinary Collaboration: An Emerging Cognitive Science* (pp. 3-21). Mahwah, NJ: Lawrence Erlbaum Associates.

Committee on Science, Education, and Public Policy. (2004). *Facilitating Interdisciplinary Research*. Washington, D.C.: National Academy of Sciences, National Academy of Engineering, Institute of Medicine.

DERRY, S. J., & SCHUNN, C. D. (2005). Interdisciplinarity: A Beautiful but Dangerous Beast. In S. J. DERRY, C. D. SCHUNN & M. A. GERNSBACHER (Eds.), *Interdisciplinary*

Collaboration: An Emerging Cognitive Science (pp. xiii-xx). Mahwah, NJ: Lawrence Erlbaum Associates.

GROOM, N. W. (2007). Phra*seology and epistemology in humanities writing: a corpus-driven study*. Unpublished PhD Thesis. Birmingham: University of Birmingham.

HALLIDAY, M. A. K. (1994). *An Introduction to Functional Grammar* (2 ed.). London: Edward Arnold.

MATHIAS, P. (2009). *A Role for Applied Linguistics in Interdisciplinary and Interprofessional Cooperation*. Unpublished MA Dissertation. Birmingham: University of Birmingham.

OAKEY, D. J. (2002). Lexical Phrases for Teaching Academic Writing in English: Corpus Evidence. In S. NUCCORINI (Ed.), *Phrases and Phraseology - Data and Descriptions* (pp. 85-105). Bern: Peter Lang.

OAKEY, D. J. (2009). Fixed Collocational Patterns in Isolexical and Isotextual versions of a Corpus. In P. BAKER (Ed.), *Contemporary Corpus Linguistics* (pp. 142-160). London: Continuum.

LEXCONN: a French Lexicon of Discourse Connectives

Charlotte Roze[1] Laurence Danlos[1] Philippe Muller[2]
(1) Université Paris 7, Alpage
(2) Université Toulouse, IRIT & INRIA, Alpage
charlotte.roze@linguist.jussieu.fr, laurence.danlos@linguist.jussieu.fr,
muller@irit.fr

Abstract. With respect to discourse organisation, the most basic way of signalling the speaker's or writer's intentions is to use explicit lexical markers: so-called discourse markers or discourse connectives. While a lexicon of discourse connectives associated with the relations they express can be very useful for researchers, especially in Natural Language Processing, few projects aim at collecting them exhaustively, and only in a small number of languages.

We present LEXCONN, a French lexicon of 328 discourse connectives, collected with their syntactic categories and the discourse relations they convey, and the methodology followed to build this resource. The lexicon has been constructed manually, applying systematic connective and relation identification criteria, using the Frantext corpus as empirical support. Each connective has been associated to a relation within the framework of Segmented Discourse Representation Theory. We make a case for a few refinements in the theory, based on cases where no existing relation seemed to match a connective's usage.

Keywords. discourse connectives, discourse relations, lexicon, ambiguity

1 Introduction

With respect to discourse organisation, the most basic way of signalling the speaker's or writer's intentions is to use explicit lexical markers: so-called discourse markers or discourse connectives. Used to express functional relations between parts of discourse, these items can be used at the sentential level or at the level of larger textual units.

We will focus here on the basic inter-sentential level: what is expressed as a whole by two sentences in a coherent discourse. This can be recursively extended to cover an entire discourse when the same relations are applied to sets of sentences. Discourse connectives explicitly signal the presence of a discourse relation between two discourse units and more generally, they contribute to discourse coherence and mark discourse structure, at least the basic organisation mentioned in Spooren and Sanders (2008): causality, sequence, grouping, contrast.

From the reader's point of view they help to disambiguate discourses whose interpretations would be vaguer without them. For example, in (1a), two interpretations are possible:[1] either Peter can find his own way home because he is not stupid (relation *Result*), or the fact that Peter can find his own way home proves he is not stupid (relation *Evidence*). We can see in (1b) and (1c) that the connectives (which are italicized) forces one of the two interpretations.

[1] This example comes from (Wilson and Sperber, 1993).

(1) Peter is not stupid.
  a. He can find his own way home.
  b. *So* he can find is own way home.
  c. *After all*, he can find is own way home.

A lexicon of discourse connectives associated with the relations they express can be very useful for researchers in Natural Langage Processing, who aim at produce automatic discourse analysis for French. Connectives can help to select the right relation between two discourse units, as they do for speakers. Very few studies or projects aim at collecting them exhaustively, and only in a small number of languages. We will detail the gathering of such a resource for French, LEXCONN,[2] and the methodology followed. The set of functional and rhetorical relations targeted by this study is taken *a priori* from Segmented Discourse Representation Theory (Asher and Lascarides, 2003), and we will evaluate how good a fit the theory is with respect to the set of connectives under investigation.

In LEXCONN we list 328 discourse connectives, collected with their syntactic categories and the discourse relations they express. Such a resource already exists for English (Knott, 1996), Spanish (Alonso et al., 2002) and German (Stede and Umbach, 1998), but LEXCONN is the first one for French. The lexicon aims at being exhaustive. It has been constructed manually, applying systematic connective identification criteria, associating a SDRT relation, and the type (coordinating or subordinating) of this relation with each connective. We used the FRANTEXT[3] corpus as a source of examples.

The rest of the paper is organised as follows. In Section 2, we present the theoretical background of this work (SDRT) and introduce the terminology we adopt about discourse connectives. In Section 3, we detail the methodology for building the lexicon and present syntactic, semantic and discursive criteria for identification of connectives. In Section 4, we describe the second stage of our work: associating discourse relations with discourse connectives. In Section 5, we present some problematic cases for SDRT when trying to associate relations with connectives.

2 Preliminaries

Our work is in line with SDRT (Asher and Lascarides, 2003), who inherits from the Discourse Representation Theory or DRT (Kamp, 1981) and discourse analysis (Grosz and Sidner, 1986; Mann and Thompson, 1988). SDRT aims at representing discourse coherence and discourse structure. The construction of SDRS (Segmented Discourse Structures) mainly rests on the distinction between coordinating relations (like *Narration* and *Result*) and subordinating relations (like *Elaboration* and *Explanation*). This distinction allows for the definition of some important principles of the theory, such as the *Right Frontier Constraint* (RFC). According to this constraint, in the course of building an SDRS, the only available sites for attachement of new information are the last segment of the discourse context and the segments which structurally dominate it.

Following Danlos (2009), we use the following terminology. The clause where a connective appears is called its "host clause". A discourse connective/relation has two arguments which

[2] The data base is available at www.linguist.univ-paris-diderot.fr/~croze/.

[3] FRANTEXT is a textual base of French litterature. It is available at www.frantext.fr.

are the semantic representations of two discourse segments called “host segment” and “mate segment”. The host segment of a connective is identical to or starts at its host clause. The mate segment is governed by constraints described in Section 3.1.

3 Building a Lexicon of Connectives

The first step of our methodology was to gather a corpus of discourse connectives candidates (about 600). To do that, we used various corpora of conjunctions of subordination and prepositions given by Eric Laporte and Benoît Sagot, the list of French discourse markers of the ANNODIS project[4] and the translated corpus of English discourse connectives built by Knott (1996).

In the database, we associate a syntactic category with each connective, which can differ a little from traditional ones: conjunction of coordination (`cco`) for connectives like *et* (*and*), *ou* (*or*) and *mais* (*but*), which are always at the beginning of their host clause, and whose mate segment is always on the left; conjunction of subordination (`csu`) for connectives like *parce que* (*because*), *même si* (*even though*) and *tandis que* (*whereas*), which are always at the beginning of their host clause, and whose mate segment can be anteposed, postposed, or internal;[5] preposition (`prep`) for the reduced forms of conjunctions of subordination when the host clause is an infinitive VP, like *afin de* (*in order to*), *pour* (*for*) and *avant de* (*before*);[6] adverb (`adv`) for connectives like *donc* (*so*), *néanmoins* (*nevertheless*) and *en tout cas* (*in any case*), which can appear in various positions in their host clause, and whose mate segment is always on the left.[7]

After gathering a corpus of candidate connectives, we have applied various criteria for the identification of connectives. In Section 3.1, we present some syntactic and semantic criteria we used for identification of connectives, and in Section 3.2, some discursive ones.

3.1 Syntactic and Semantic Criteria

The criteria we present in this Section concern three properties of discourse connectives: they are not integrated to propositional content (cleft criterion), they cannot be referential expressions (substitutability criterion), and their meaning is not compositional (compositionality criterion).

Cleft Criterion Discourse connectives cannot be focused in cleft constructions.

According to Riegel et al. (2004), the items which can be focused in cleft constructions have one of the following functions: subject, object, or adverbial. These items are inside the predicative structure. Jayez and Rossari (1996) distinguish the connectives which are integrated to the predicative structure (and which can be focused in cleft constructions) from the other ones. For example, they claim that *à ce moment-là* in (2a) is a temporal connective which can be focused

[4] ANNODIS is a project of French discourse annotation (Péry-Woodley et al., 2009).

[5] However, for some conjunctions of subordination like *comme*, the mate segment is always anteposed. For others, the mate segment can be anteposed or internal. These informations are marked in LEXCONN.

[6] There exists a few `prep` which are not linked with `csu`, e.g. *quitte à*, *quant à*.

[7] We consider as adverbs some NPs which are not introduced by a preposition, like *la preuve*, *résultat*.

in a cleft construction, see (2b). On the other hand, Bras (2008) claims that *à ce moment-là* in (2a) is not a connective, but a temporal cue: it only temporally locates events, and doesn't play any role at the discourse level. We agree with Bras contra Jayez and Rossari: *à ce moment-là* has a non-discourse usage in (2a), where it refers to the temporal location of an eventuality, while it has a discourse usage in (3a) where it cannot be clefted, see (3b). Moreover, it is referential in (3a) but not so in (3b), which goes along with the next criterion.

(2) *Il a commencé à pleuvoir.* 'It started raining.'
  a. A ce moment-là*, Marie est arrivée.* '*At that moment*, Mary arrived.'
  b. C'est à ce moment-là que *Marie est arrivée.*

(3) *Tu as l'air de penser qu'elle n'est pas honnête.* 'You seem to think she is not honest.'
  a. A ce moment-là*, ne lui raconte rien.* '*So* don't tell her anything.'
  b. # C'est à ce moment-là que *ne lui raconte rien.*

Substitutability Criterion Discourse connectives cannot be substituted by an entity (person, event, discourse unit) of the context.

Knott (1996) considers as discourse connectives some phrases like *because of this*. He keeps phrases which contain propositional anaphora in his corpus, which can be substituted by entities of the discourse context. On the contrary, we don't retain this type of phrases in LEXCONN.

To illustrate the Substitutability Criterion, consider *après ça* in (4b) and *à part ça* in (6b). On the one hand, in (4b), *ça* refers to the segment in (4a), as shown by the acceptability of (5). On the other hand, *ça* in (6b) does not refer to the segment in (6a), as shown by the inacceptability of (7). The Substitutability Criterion tells us that *après ça* is not a connective, while *à part ça* remains in the corpus of candidate connectives.

(4) a. *Bruno est allé en Argentine.* 'Bruno went to Argentina.'
  b. Après ça*, il est parti au Pérou.* '*After that*, he moved to Peru.'

(5) Après *[qu'il est allé en Argentine], Bruno est parti au Pérou.*

(6) a. *Hier soir j'ai croisé Pierre dans une boîte de nuit.* 'Last night I saw Peter in a nightclub.'
  b. A part ça *il nous dit tout le temps qu'il est fatigué.* '*Though* he always says he is tired.'

(7) # A part *[qu'hier soir je l'ai croisé dans une boîte de nuit], Pierre nous dit tout le temps qu'il est fatigué.*

Compositionality Criterion Discourse connectives are invariable.[8]

Various studies (Molinier, 2003; Cojocariu and Rossari, 2008; Nakamura, 2009) aim at showing the connecting role played by adverbials like *à ce propos* and *la preuve*, which contain (predicative) nouns. It seems that the emergence of a discursive role for these adverbials is correlated with a process of fixation. For example, the determiners and the numbers of *la preuve* and *à*

[8] Connectives cannot undergo internal modification, but some of them can be externally modified by adverbials, such as *probablement* or *certainement* for *parce que*.

ce propos (in their discourse usages) have become invariable (# *les preuves*, # *à ces propos*). These studies inspired our Compositionality Criterion: nouns contained in connectives cannot be modified by an adjective, their numbers and their determiners are invariable. This criterion allows us to retain some candidates like *en tout cas* and *résultat*: *en tout cas* in (8a) cannot be modified by an adjective in (8b), and *résultat* in (9a) is invariable, see (9b).

(8) *Je ne sais plus s'il y avait vraiment de la neige, ce Noël-là.* 'I don't know if there really was snow, that Christmas.'

  a. En tout cas*, dans mon souvenir, je la vois tomber...*[9] '*In any case*, I remember seeing it falling...'

  b. # En tout cas envisagé / possible*, dans mon souvenir, je la vois tomber...*

(9) *Pierre n'a pas réussi à dormir cette nuit.* 'Peter couldn't spleep last night.'

  a. Résultat*, il était en retard aujourd'hui.* '*Thus*, he was late today.'

  b. # Le résultat / Les résultats*, il était en retard ce matin.*

3.2 Discursive Criteria

The criteria we present in this Section only make use of discourse notions. They were applied after syntactic and semantic criteria, and helped identifying discourse relations conveyed by connectives.

Contextual Criterion If the discourse $D = c\ clause$ is coherent without other discourse context, then c is not a discourse connective.

The Contextual Criterion is the only test Knott (1996) used to build a list of English connectives. This test is insufficient to discard adverbials like *le lendemain* or *un peu plus loin*, which express temporal or spatial information. However we used Knott's test to discard some candidates.

Forced Relation Criterion Let D_a and D_b be coherent discourses with $D_a = seg_1\ seg_2$ and $D_b = seg_1\ c\ seg_2$, R_a the discourse relation which holds between seg_1 and seg_2 in D_a, and R_b the relation which holds in D_b. If $R_a \neq R_b$ then c is a discourse connective.

Consider (10a) and (10b) which differ by the presence of *malheureusement* in (10b). The segment in (10a) is an *Explanation* of the first segment (Mark will camp this summer), whereas the segment in (10b) is in a *Contrast* relation with the first segment (maybe Mark will not camp this summer). This is evidence that *malheureusement* is a connective. On the other hand, consider (10c) and (10d) which differ by the presence of *évidemment* in (10d). The presence of this adverbial doesn't change the discourse relation, which is *Result* in both cases. More generally, we found no example where the presence of this adverb changes the relations involved. This is evidence that *évidemment* is not a connective.

(10) *Marc veut faire du camping cet été.* 'Mark wants to camp this summer.'

  a. *Il n'a pas beaucoup d'argent.* 'He does not have much money.'

[9] Patrick Modiano, *Un pedigree*, 2005, p. 94.

b. Malheureusement *il n'a pas beaucoup d'argent.* '*Unfortunately* he does not have much money.'

c. *Il faut qu'il économise de l'argent.* 'He must save up money.'

d. Evidemment*, il faut qu'il économise de l'argent.* '*Of course*, he must save up money.'

Coherence Criterion If seg_1 seg_2 is incoherent and seg_1 c seg_2 is coherent, then c is a discourse connective.

Beaulieu-Masson (2002) gives a study of connectives like *à propos*, *à ce propos* and *au fait*, which force discourse coherence. For example, in (11), the presence of *à propos* helps linking the segment in (11b) to the segment in (11a). Without it, the discourse would be incoherent. The Coherence Criterion is inspired from this study. It can be used for various connectives. For example, *ceci dit* in (12a) is a discourse connective (which mark the relation *Opposition*), because if it is deleted, the discourse becomes incoherent, see (12b).

(11) a. *Boris, Je prends des gouttes pour stimuler mon appétit, mais les résultats sont lents, très lents.* 'Boris, I take drops to stimulate my appetite, but the results are slow, very slow.

b. A propos*, vers quel moment crois-tu que tu pourras venir ?*[10] '*By the way*, when can you come ?'

(12) *Ce serait vraiment utile pour nous d'aller à cette réunion.* 'It would be really useful for us to go to this meeting.'

a. Ceci dit*, on peut s'en passer.* '*But* we can do without it.'

b. # *On peut s'en passer.* 'We can do without it.'

After we applied these criteria, 328 candidates were kept as connectives.[11]

4 Associating Relations with Connectives

After building the list of French discourse connectives, we tried for each connective to determine which discourse relation(s) it expresses, observing the contexts where it appears in discourses from the FRANTEXT corpus. To do this, we used a set of 15 discourse relations defined in SDRT, which are of various kinds: temporal (*Narration*, *Background* (*backward* or *forward*), *Flashback*), causal (*Result*, *Explanation*, *Goal*), structural (*Parallel*, *Contrast*, *Elaboration*, *Continuation*), logical (*Alternation*, *Consequence*), metatalk (*Result**, *Explanation**). Each relation is typed (coordinating or subordinating), and has semantic effects.

4.1 Tests for Relations Identification

In order to identify the discourse relation conveyed by a connective, we tried to use the following clues.

[10] Lydia Flem, *Lettres d'amour en héritage*, 2006, p. 127.

[11] The list of discourse markers from ANNODIS project contains about 60 connectives.

Attachment Test This test helps to determine the type of the relation (Asher and Vieu, 2005). As we said in Section 2, in SDRT, relations are either coordinating or subordinating. This distinction is essential for the RFC: if the relation between two discourse segments (π_1) and (π_2) is subordinating, a third segment (π_3) can be attached to (π_1), whereas if it is coordinating, (π_3) cannot be attached to (π_1), because (π_1) is no longer available for attachment. We used this test to identify the type of relation expressed by connectives.

Substitution Test If two connectives are substituable for each other in most of the discourse contexts they appear in, e.g. the discourse interpretation is unchanged, they probably express the same discourse relation. This test is inspired from Knott (1996). However, given that our goal is not to build a taxonomy of connectives/discourse relations we did not use more subtle relationships than contingent substituability (such as synonymy, hyponymy or hyperonymy).

For example, the Substitution Test tells us that *dès lors que*, *puisque* and *étant donné que* have one discourse usage in common: in (13), they are substituable for each other without changing the discourse interpretation (they express $Explanation*$).

(13) *Brillant résultat de quinze ans de diplomatie gaulliste, mais résultat inévitable,* dès lors que / puisque / étant donné que *nous avons toujours placé (...) les apparences au-dessus des réalités ...*[12] 'This is the brilliant outcome of fifteen years of Gaullist diplomacy, but this is inevitable, *given that* we always preferred appearances to reality.'

Semantics Effects In SDRT, discourse relations have semantic effects. Some relations (such as $Background$, $Explanation$ and $Flashback$) set temporal constraints on the eventualities they link. For example, $Flashback(\alpha, \beta)$ implies a temporal precedence between e_α and e_β.[13] Relations such as $Result$ and $Explanation$ can also establish causal relationships between eventualities. For instance, $Result(\alpha, \beta)$ implies a causal link between e_α and e_β.

4.2 Ambiguity

The database contains 328 connectives, and 428 usages of connectives: connectives are ambiguous. We describe here two types of ambiguity.

Some connectives can establish more than one discourse relation. For instance, *si* has a conditional usage (see (14)), in which its mate segment can be anteposed, postposed or internal. It also has a concessive usage (see (15)), in which its mate segment can only be anteposed. In the same way, the adverb *aussi* expresses $Result$ when it is in initial position of its host clause and $Parallel$ when it is not in initial position.

(14) Si *je ne reçois pas très vite de l'aide, nous courons au désastre.*[14] 'If nobody comes to my help very soon, we're doomed.'

(15) *Quand j'étais un jeune garçon, j'ai manié indéfiniment les vieux fascicules de cette revue.* Si *j'étais trop jeune pour les bien comprendre, j'en recevais toutes sortes de rêves...*[15] 'When I was a boy, I handled old issues of this magazine endlessly. If I was too young to understand them, I drew all kinds of dreams from them...'

[12] Pierre Mendès-France, *Oeuvres complètes. 6. Une vision du monde.*, 1974-1982, 1990, p. 133

[13] e_α and e_β are the eventualities described in the segments α and β.

[14] Patrick Rambaud, *La Bataille*, 1997, p. 228, CHAPITRE V, Seconde journée.

[15] Inspired from: Maurice Barrès, *Mes Cahiers - Tome 14 : 1922-1923*, 1923, p. 163, 46ème cahier.

In LEXCONN, such informations about the position of the mate segment of subordinating conjunctions and the position of adverbs in their host clause are encoded by specific attributes/features (`position-sub` and `position-adv`). However, for many ambiguous connectives, the usage cannot be selected by surface clues like the connective's position or the mate segment's position and depends more on discourse content.

Some other connectives such as *et* (*and*) present a second type of ambiguity : they have discourse and non-discourse usages. These non-discourse usages are frequent for adverbials and are not represented in LEXCONN. However we kept in the lexicon non-discourse usages for connectives like *à ce moment-là* ($Result*$) and *en même temps* ($Opposition$), which often express strictly temporal relations (e.g. temporal simultaneity).

We now give quantitative data about ambiguous connectives:[16] 73 connectives (23,7%) have more than one discourse usage and 14 connectives (4,2%) have discourse and temporal usages. Concerning ambiguity between discourse usages, two cases must be distinguished: the case where a connective establish discourse relations of the same type (coordinating or subordinating) and the case where a connective establish relations with different types. The first case seems less problematic than the second in an NLP perspective, because it doesn't implies structural ambiguity. Only 6,2% from the total number of connectives are in the second case.

4.3 Relations Frequency

We cannot yet know the frequency of each discourse connective in terms of occurrences in a real corpus (this work has to be done using LEXCONN and the ANNODIS corpus), but we now can give the frequency of each discourse relation in terms of number of connectives. These frequencies are given in Table 1.[17] Some of the relations are defined in SDRT and listed above, but some of them are not and are detailed in Section 5.

About 28% connectives are "contrastive" ones, e.g. they express either *Contrast* (formal contrast) or *Opposition* (violation of expectation) or *Concession* (these relations are grouped together in ANNODIS corpus). What we can say is that there are many ways of expressing contrastive relations, maybe because they are difficult to express without a discourse connective. On the contrary, *Elaboration* has a low frequency in terms of connectives.

5 Problematic Cases for SDRT

This stage led to the following result about discourse relations: some discourse connectives appear in contexts where no relation defined in SDRT can hold. In other words, although this work is in line with SDRT, the set of discourse relations defined in the theory is insufficient for describing the contributions to discourse interpretation of all French discourse connectives. Two cases must be distinguished. First, the case where we can introduce relations that are not defined in SDRT. These relations are generally defined in Rhetorical Structure Theory (Mann and Thompson, 1988). Second, the case for which it seems impossible to associate any relation to a discourse connective.

[16] We do not consider connectives marked as "unknown" in the counts.

[17] Notice that we distinguish several usages for some connectives.

Relation	Number	Percentage
Opposition	41	9,5
Result	35	8,1
Concession	32	7,4
Continuation	32	7,4
Explanation	28	6,5
Goal	25	5,8
Condition	25	5,8
$Explanation*$	24	5,6
Narration	23	5,4
Unknown	21	4,9
Contrast	17	4,0
$Background_b$	15	3,5
$Temporal_{location}$	14	3,3
Parallel	13	3,0
Elaboration	11	2,6
$Result*$	11	2,6
Summary	11	2,6
Flashback	10	2,4
Detachment	9	2,1
Alternation	9	2,1
Consequence	7	1,6
$Background_f$	7	1,6
Evidence	7	1,6
Rephrasing	6	1,4
Digression	6	1,4
Total	428	100%

Table 1: Relations frequencies: number and percentage of connectives.

5.1 Introducing New Relations in SDRT

We introduced six relations in LEXCONN which are not defined in SDRT. These relations are: *Concession* (même si, bien que), *Opposition* (cependant, malgré tout), *Summary* (en gros, globalement), *Detachment* (quoi qu'il en soit, de toute manière), *Digression* (à propos, au fait), and *Rephrasing* (enfin, tout au moins). These relations were introduced because no relation defined in SDRT can represent the contributions to discourse interpretation of some connectives, which can be grouped together with respect to the contexts where they appear.

For example, connectives like *bien que* or *même si* are considered in ANNODIS as markers of the coordinating relation *Contrast*. However, they express a subordinating relation, as shown in (16): the segments (π_1) and (π_3) are linked by the relation *Result*, therefore the relation between (π_1) and (π_2) is necessarily subordinating. The discourse structure associated with (16) is shown in Figure 1.

(16) a. *Pierre m'a aidé à repeindre la chambre* 'Peter helped me repaint the bedroom' (π_1)

b. bien qu'*il ait beaucoup de boulot en ce moment*. '*even though* he has a lot of work at this time.' (π_2)

c. Du coup*, c'est déjà terminé !* '*Thus* it is already over.' (π_3)

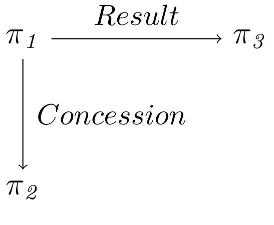


Figure 1: Graph Representation for (16)

In addition, these connectives link segments which don't have necessarily similar semantic structures, while *Contrast* must link segments with some structural similarities. In conclusion, *bien* *que* and *même si* cannot express the coordinating relation *Contrast*: they express the subordinating relation *Concession* (which is defined in RST) that we introduced in LEXCONN.

5.2 Unknown Relations

For 21 connectives (about 6%), the associated discourse relation in LEXCONN is `unknown`. Among these connectives, there are adverbs (*en fait*, *au moins*), conjunctions of subordination (*avant même que*, *à mesure que*), and prepositions (*quant à*, *quitte à*). Each connective associated with `unknown` verifies the criteria we presented in Section 3, but any possible relation is insufficient for describing the semantics of the connective.

For example, *à mesure que*, whose meaning is non-compositional, as shown by the inacceptability of (17b), and which doesn't contain a referential expression, as shown in (17c), is a connective. However, whatever relation we try to associate with it (*Simultaneity*, *Explanation*, or even *Parallel*), some semantic information is lost, i.e. the fact that there is a simultaneous temporal progression between the two events involved. As a consequence, *à mesure que* is associated with `unknown`.

(17) *Tes digressions s'allongeaient* 'Your digressions got longer and longer'
   a. à mesure que *tu finissais les alcools de ta mère.*[18] '*as and when* you finished your mother's alcohols.'
   b. # à la mesure que *tu finissais les alcools de ta mère*.
   c. # à cette mesure-là.

6 Conclusion

Building a French lexicon of discourse connectives brought several results. It involved a systematic methodology to identify discourse connectives and associate discourse relations to them, resting on various studies about connectives and corpus-collected examples. In addition, it shows which connectives remain to be studied in detail (especially connectives whose function is "unknown" so far). A statistical analysis of the resulting lexicon allowed us to quantify several things, such as the importance of the various discourse relations in terms of the number of connectives associated with them, and a count of ambiguous connectives.

Despite these results, LEXCONN has to be improved: some information has to be added. For example, some information about ambiguity between discourse and non-discourse usage has to be introduced. This improvement will be possible with other linguistic analysis, but also with automatic analysis on ANNODIS corpus: we could examine the link between position in the host clause and discursive/non-discursive role for adverbials.

However, LEXCONN already constitute a precious resource for NLP. It might help for discourse markers annotation in ANNODIS, in which connectives are not yet marked. A statistical analysis of the connectives on corpus can also be useful, for example concerning connective's frequency. Such analysis could help answering the following question: are ambiguous connectives the most frequent ones?

[18] Edouard Levé, *Suicide*, 2008, p. 29.

References

Laura Alonso, Irene Castellón, and Lluís Padró. Lexicón computacional de marcadores del discurso. *SEPLN, XVIII Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 2002.

Nicholas Asher and Alex Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

Nicholas Asher and Laure Vieu. Subordinating and coordinating discourse relations. *Lingua, Elsevier*, 115(4):591–610, 2005.

Anne Beaulieu-Masson. Quels marqueurs pour parasiter le discours ? *Cahiers de Linguistique Française*, 24:45–71, 2002.

Myriam Bras. *Entre relations temporelles et relations de discours*. Dossier d'HDR, Université de Toulouse le Mirail, 2008.

Corina Cojocariu and Corinne Rossari. Constructions of the type *la cause/la raison/la preuve* + utterance: grammaticalization, pragmaticalization, or something else? *Journal of pragmatics*, 40:1435–1454, 2008.

Laurence Danlos. D-STAG: a formalism for discourse analysis based on SDRT and using synchronous TAG. In *Proceedings of the 14th Conference on Formal Grammar (FG'09)*, pages 1–20, 2009.

Barbara Grosz and Candace Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204, 1986.

Jacques Jayez and Corinne Rossari. *Donc* et les consécutifs, des systèmes de contraintes différentiels. *Linguisticæ Investigationes*, XX:117–143, 1996.

Hans Kamp. Evénements, représentations discursives et référence temporelle. *Langages*, 64: 34–64, 1981.

Alistair Knott. *A Data-Driven Methodology for Motivating a Set of Coherence Relations*. PhD thesis, Department of Artificial Intelligence, University of Edinburgh, 1996.

William Mann and Sandra Thompson. Rhetorical structure theory: Towards a functional theory of text organization. *Text*, 8:243–281, 1988.

Christian Molinier. Connecteurs et marqueurs énonciatifs : Les compléments figés formés à partir du nom *propos*. In *Actes du Colloque Grammaires et Lexiques Comparés*, volume 26, pages 15–31. Conenna, Mirella and Laporte, Éric, 2003.

Takuya Nakamura. Observations sur la prédication : prédicat verbal, prédicat nominal avec verbe support et prédicat nominal sans verbe support. In *Actes du Colloque International Supports et prédicats non verbaux dans les langues du monde*, Paris, France, 2009.

Marie-Paule Péry-Woodley, Nicholas Asher, P. Enjalbert, Farah Benamara, Myriam Bras, Cécile Fabre, Stéphane Ferrari, Lydia-Mai Ho-Dac, Anne Le Draoulec, Yann Mathet, Philippe Muller, Laurent Prévot, Josette Rebeyrolle, Ludovic Tanguy, Marianne Vergez Couret, Laure

Vieu, and Antoine Widlöcher. ANNODIS : une approche outillée de l'annotation de structures discursives (poster). In *Traitement Automatique des Langues Naturelles (TALN 2009)*, Senlis, France, 2009.

Martin Riegel, René Rioul, and Jean-Christophe Pellat. *Grammaire méthodique du français*. Presses universitaires de France, Paris, France, 2004.

Wilbert Spooren and Ted Sanders. The acquisition order of coherence relations: On cognitive complexity in discourse. *Journal of Pragmatics*, 40:2003–2026, 2008.

Manfred Stede and Carla Umbach. Dimlex: A lexicon of discourse markers for text generation and understanding. In *In Proceedings of the Joint 36th Meeting of the ACL and the 17th Meeting of COLING*, pages 1238–1242, 1998.

Deirdre Wilson and Dan Sperber. Linguistic form and relevance. *Lingua*, 90:1–25, 1993.

Text Organization: Identifying and Measuring the Strength of Arguments in Procedural Text

Lionel Fontan, Patrick Saint-Dizier
IRIT, 118 route de Narbonne
31062 Toulouse Cedex France
fontan@univ-tlse2.fr , stdizier@irit.fr

1 Motivation and Aims

Argumentation (e.g. (Amgoud et al 2001, Moeschler 1985)) and, in particular, persuasive argumentation is a process frequently encountered in several types of texts where the challenge is to convince the reader to adhere to a certain point of view. Arguments come with forms of emphasis which give them more strength than normally expected, or, conversely, they may come with forms of irony or of depreciation, which influence the reader's perception of the facts associated with the arguments. They are realized by a variety of signals whose study is of much interest, in particular in Web documents. Signals may be terms like adverbs of intensity as well as icons, font sizes, etc. Depending on the author, the target audience and the domain at stake, the type of signal can vary greatly.

Persuation appears in different types of texts with similar objectives but with slightly different linguistic, layout and typographic forms. This is, for example, the case in legal text analysis (Moens et alii., 2007). The situation of procedural texts, although ranging over a large set domains, seems to be simpler in terms of linguistic forms and underlying interpretation(s). One of the reasons is that procedural texts are basically action-oriented, and, therefore, the number of inferences that the user may have to do is limited as much as possible. Nevertheless, there are crucial problems associated with argumentation and persuation which are typical of procedural texts: arguments, in particular warnings, implicitly indicate that some actions are difficult to realize, and that there is a risk of failure (Dautriche et al. 2009). In terms of Action Theory, this is an interesting way to measure the complexity of a procedure and the chances to succeed, or the risks to fail.

Most of these aspects are made very explicit in procedural texts, whatever their style, by means of (1) very explicit, recurrent and domain-independent linguistic marks, (2) relatively clearly identified and recurrent icons, punctuation forms and typographic forms and (3) a global text architecture and possibly annotations, as in technical documents (maintenance manuals is a typical illustration). Obviously the styles and the related signals are very diverse depending on the domain, the authors and even more crucially, the target readers.

The challenge in procedural texts is to convince the reader that the procedure which is proposed for reaching a certain goal (concrete as in do-it yourself texts, or more abstract as in social relation texts) is among the bests, that the user gets excellent and adequate help, hints and advice while following the procedure and that results are guaranteed, modulo some precautions (e.g. caring about warnings, reading and considering advice, carefully realizing instructions in the order they are given, etc.). It is a way of 'selling' the procedure, in comparison with other procedures describing the same task (since the web abounds in procedures, often quite different in form and contents, for realizing a certain task).

A second type of underlying objective is to make sure that the reader, when realizing the procedure, will effectively strictly and fully realize the instructions as they are given, while indicating him that otherwise he may undergo problems. In procedural texts, this is essentially realized by means of advice and warnings. It seems that these tow forms of argumentation in procedural texts follow a small number of quite standard schemas (Walton et ali., 2008). Finally, a third register in persuation, positively oriented, consists in supporting the reader when the task is complex, long or risky.

In conjunction with arguments, procedural texts abound in persuasive forms of various kinds. These forms are made visible via by a variety of marks, essentially linguistic, but also typographic, iconic or even possibly by means of images. At a global level, the presence of a number of advice and warnings in a text, is, by itself a form of persuation based on an implicit perception by the user that the text has received an in-depth elaboration and results from a long experience. Besides persuasive arguments, we observed a variety of explanation forms which have a certain implicit persuasive impact, such as reformulations, hints, definitions, etc. Besides persuation, at a theoretical level, it is of much interest to define a formal model of procedurality in terms of Action Theory (Dautriche et al. 2009). Within procedures, a number of persuasive forms also introduce some form of comfort for the user, so that he can work safely and without too much stress and worries.

This work is part of a larger project, Anonymous dedicated to procedural text processing, and various operations on these texts such as enrichment, incoherence analysis, fusion, etc. A framework around Action Theory has been developed to give a formal semantics to our analysis.

2 The explanation structure in procedural texts

2.1 A global view of the explanation structure

We first constructed a quite large corpus of texts oriented towards action (about 1700 texts in French from a large number of web sites) from several domains. These texts which are, roughly, procedural texts, are quite diverse in style and complexity, from cooking, do it yourself, gardening, equipment maintenance, to social relations, health, and didactics. Those texts are in general not very long, ranging from half a page to 4 pages.

From this corpus, we established a classification of the different forms explanations may take in such a type of text (Fontan et al 2008). The main structures we identified are facilitation and argumentation structures. These structures are organized as follows:

- **facilitation structures**, which are rhetorical in essence (Kosseim et al. 2000, Van der Linden 1993), correspond to *How to do X ?* questions, these include two subcategories:
  (1) user help, with: hints, evaluations and encouragements and
  (2) controls on instruction realization, with two cases:

  (2.1) controls on actions: guidance, focusing, expected result and elaboration and
  (2.2) controls on user interpretations: definitions, reformulations, illustrations and also elaborations.

- **argumentation structures**, corresponding to *why do X ?* questions. These have either:
  (1) a positive orientation with the author involvement (promises) or not (advice and justifications) or
  (2) a negative orientation with the author involvement (threats) or not (warnings).

In procedural texts, we essentially observed advice and warnings since there is seldom any involvement from the author.

User help structures aim at making the user more comfortable with the current document: the way hints (*prefer a sharp knife*) and encouragements (*at this stage you've done the difficult part*) are termed and are perceived by the reader is a crucial step in the persuation process. Evaluations are in general accurate and positively oriented, guiding the user and preventing him from any questioning and discouragements (*now your sauce must look yellow, if not add more flour*). User guidance and controls on user interpretation provide the necessary assistance (possibly user parameterized, depending e.g. on how much interactions the user wishes, the type of help it requires, etc.) to guarantee a certain success, in particular when the procedure is difficult or long, with several subparts. This contributes to a feeling of control and safety w.r.t. actions being realized.

2.2 Arguments in Explanation Structures

Arguments in procedural texts serve very different purposes. They make explicit the risks that the user may undergo if he does not follow the instructions, its responsability is clearly made explicit and his role is more active. In terms of persuation, the strength of the arguments and the illocutionary force of the statements aim at convincing the reader of the reality and the importance of the risks, in the case of warnings, or of the gains in the case of advice.

It is important to note that all these aspects do not operate in isolation, but they all contribute to the success of the procedure realization. For example, well designed hints will convince the reader that the document is of high quality and that, therefore, warnings should be taken seriously.

The most appropriate structure in which arguments appear is neither the instruction nor the whole text, but an intermediate structure that has some autonomy and coherence that we call an instructional compound. It is basically organized around a few kernel instructions and is modified by a number of structures sucha s conditionals, goal expressions, and a number of rhetorical segments (elaboration, illustration, reformulation, etc.) among which, most notably, arguments. An example, using the square bracket notation, of arguments within an instructional compound is:

$[_{instructional\ compound}$
$[_{Goal}$ To clean leather armchairs,]
  $[_{argument:advice}$
$[_{instruction}$ choose specialized products dedicated to furniture,
  $[_{instruction}$ and prefer them colorless]],
  $[_{support}$ they will play a protection role, add beauty, and repair some small damages.]]]

We have here an argument of type advice which is composed of two instructions (or conclusions) and a conjunction of three supports which motivate these two instructions.

The explanation structure is realized by language expressions, characterized by dedicated linguistic marks typical of help statements, reformulations, etc. The typography is also an important factor via the ease of readability it introduces and also by the professionalism it suggests. The major elements are given below. Obviously, the impact to the layout in general is crucial but it is very difficult to formally measure.

Our goal is to identify and categorize most of these marks, and then to a priori sort them on various scales related to persuation strength, so that, ultimately, the parameters of persuation can be measured on a given procedural text, instruction by instruction. It is then also crucial to evaluate how these elements are perceived and interpreted by a variety of users. It is obviously difficult to derive a formal model due to the subjectivity of the measures (Grosz et al. 1986): in this short document, we focus on argument strength identification.

3 Processing arguments

We present here the form warnings and advice take in language expressions in procedural texts. This is the basic mark that gives the interprepation of the statement. Besides, additional marks, such as icons, punctuation and typography, reinforce or weaken the strength of the perceived argument. This is given in the next section.

The linguistic forms given below can be implemented by means of patterns. A first implementation was carried out in Perl (Anonymous) (Delpech et al. 2008). We are now designing a much more powerful environment, dedicated to procedure processing and more generally to text semantics processing, where rules with variables and gaps can be expressed. This is a much more powerful language. Implementation is based on the Java JFLEX and JCUP technology, based on an LALR(1) automaton. An interface allows grammar rule writers to express rules and constraints in the form of context free rules associated with XML annotations. A display system, based on Navitexte will be available shortly so that results can be easily accessible and also esier to debug.

3.1 Processing warnings

Warnings are basically organized around an ’avoid expression’ combined with a proposition. The variations around the ’avoid expression’ capture the illocutionary force of the argument, ordered here by increasing force, the latter expression being very strong.

We give below, for the the three major classes we have observed, the basic pattern (between quotes) for the conclusion part of the argument (which has the form of an instruction), an example and the frequency observed in our corpus:

1. ’prevention verbs like avoid’ (NP / to VP) (*avoid hot water*), (frequency: 48%)

2. ’do not / never / ... VP(infinitive) ...’ (*never put this cloth in the sun*), (frequency: 36%)

3. 'it is essential, vital, ... to never VP(infinitive)', *it is vital to never take this medicine at the beginning of the meal*, (frequency: 6%).

Supports for warnings convey statements with a negative polarity. These are identified and delimited from various marks:

1. connectors with a negative orientation such as: *sinon, car, sous peine de, au risque de* (otherwise, under the risk of), etc. verbs expressing a consequence or verbs in the conditional form (*could damage...*),

2. negative causal expressions of the form: *in order not to, in order to avoid, etc.*

3. specific verbs such as risk verbs introducing an event (*you risk to break*). In general the embedded verb has a negative polarity.

4. very negative terms, such as: nouns: *death, disease, etc.*, adjectives, and some verbs and adverbs.

We built a lexicon of about 200 negative terms found in our corpora. While forms (1) and (2) are quite standard, those in (3) and (4) are much stronger, they appear in our corpus in about 28% of the situations. As reported in (Fontan et al. 2008), we carried out an indicative evaluation (e.g. to get improvement directions) on a corpus of 66 texts over various domains, containing 262 arguments. Those texts where manually annotated by a trained linguist, and the results were then compared with the system output. We get the following results for warnings:

conclusion recognition	support recognition	(3)	(4)
88%	91%	95%	95%

(3) conclusions well delimited (4) supports well delimited, with respect to warnings correctly identified.

3.2 Processing advice

Conclusions of type advice are essentially identified by means of two types of patterns (English glosses given here):

1. advice or preference expressions followed by an instruction. The expressions may be a verb or a more complex expression: *it is advised to, prefer, it is better to, preferable to, etc.*,

2. expression of optionality or of preference followed by an instruction: *our suggestions: ...*, or expression of optionality within the instruction (*use preferably a sharp knife*).

3. very negative terms, such as: nouns: *death, disease, etc.*, adjectives, and some verbs and adverbs.

Supports of type advice are identified on the basis of 3 distinct types of patterns:

1. 'Goal exp + (adverb) + positively oriented term'. Goal expressions are e.g.: *in order to, for*, whereas adverb includes: *better* (in French: *mieux, plus, davantage*), and 'positively oriented term' includes: nouns (*savings, perfection, gain*, etc.), adjectives (*efficient, easy, useful*, etc.), or adverbs (*well, simply*, etc.). We constructed a lexicon of positively oriented terms that contains about 50 terms. Not surprinsingly, positive terms are far less numerous than negative terms.

2. Goal expression with a positive consequence verb (*favor, encourage, save*, etc.), or a facilitation verb (*improve, optimize, facilitate, embellish, help, contribute*, etc.),

3. the goal expression in (1) and (2) above can be replaced by the verb 'to be' in the future: *it will be easier to locate your keys*.

4. very negative terms, such as: nouns: *death, disease, etc.*, adjectives, and some verbs and adverbs.

advice are related to optionality or preferences. The different marks above do not introduce a priori any strong difference in terms of persuation. It seems that if some terms look stronger than others, some informal experiments tend to indicate that it is more a matter of personal interpretation.

Similarly as above, we carried out an indicative evaluation on the same corpus of 66 texts containing 240 manually identified advice. We get the following results for advice:

conclusion recognition	support recognition	(3)	(4)	(5)
79%	84%	92%	91%	91%

(3) conclusions well delimited, (4) supports well delimited, both with respect to advice correctly identified. (5) support and conclusion correctly related.

A short example of an informally annotated arguement is given in Fig. 1 hereafter. A graphical representation using the NAVITEXTE software is given at the end of this document. We plan to use norms, as suggested in the AIF project (Chesnevar et ali. 2007) for representing argument structures.

4 Linguistic Marks of Argument Strength

Argument strength is a major parameter and concern in this type of study. Let us now review linguistic and non linguistic marks related to the 'illocutionary' force of an argument, contributing to its persuasive effect, in addition to the intrinsic force of arguments presented in the

$< procedure >< title >$ How to embellish your balcony $< /title >$
$< Prerequisites >$ 1 lattice, window boxes, etc.$< /prerequisites >$
....
$< instructional - compound >$ In order to train a plant to grow up a wall, select first a sunny area, clean the floor and make sure it is flat......
$< Argument >< Conclusion\ att = "Advice" >$ You should better let a 10 cm interval between the wall and the lattice. $< /Conclusion >$
$< Support\ att = "Advice" >$ This space will allow the air to move around, which is beneficial for the health of your plant. $< /Support >< /Argument >$ ... $< /instructional - compound >$ ......
..... $< /procedure >$

Figure 1: Extract of an annotated procedure

classifications above, essentially based on linguistic forms. These marks can be combined with the basic patterns given in the previous section. The categories given below are a priori identical for any kind of argument, positive (rewards and advice) or negative (threats or warnings). We concentrate here on those criteria that reinforce the persuasive effets, their absence could lower these effects in some cases, but this is also a matter of style.

The criteria and evaluations given below emerged from a few unformal experiments carried out on readers in our lab:

- **Number of supports**: a conclusion associated with several explicit supports seems to be stronger than if it has just one: *do not open the door when washing is ongoing*). The strength of a conclusion with no supports is quite difficult to evaluate: in a number of cases, the support is not mentioned because it is obvious for the reader and would sound odd or verbose otherwise: *do not water your plants when the temperature is below zero degrees (not mentioned: because this may 'burn' the leaves).*

- **Supports associated with some forms of rhetorical developments**. We observed, especially in large public texts, the presence of segments of texts in a rhetorical relation with the argument support (Mann et al. 1988, Van der Linden 1993). Among the most frequently encountered relations we have: exemplification, elaboration, development and reformulation: *because you risk to break the connectors which cannot then be repaired,* with here a kind of development (but such relations may be difficult to assign unambiguously).

- **Position of supports in the argument**: a left-extraposed argument is stronger than when it appears at the end of the argument. This is a general rule in pragmatics, where left extraposed elements gets higher focus, since this position is not the expected one.

- **Typography and punctuation**: we identified several marks of emphasis: capital letters, large size, italics, bold, underlined, etc. Exclamation marks are also frequent (*do not leave in a humid place!*). However, typography and punctuation mark strength is relative to their global use in the procedure. If they appear exceptionally in an instruction, then they get more strength. In general procedures, except for video game solutions and similar types of texts, are quite sober and make a very limited use of punctuation. A dedicated metrics then needs to be defined.

- **Icons and other devices**: In a number of large public documents, extra-linguistic signs such as icons are very rich and very suggestive. There are many categories such as road signs, faces, etc. Their strength is important, but quite difficult to measure. As above, a profusion of these signs lowers their impact.
- **Marks of negation**: some marks of negation are stronger than others: 'never' is stronger than 'do not', *never use X*, *do not use X* and at the lower level we have advice verbs combined with a negation *we do not advise you to use this paint*.
- **Dedicated forms**: *pay attention:, important:, advice:, etc.*, these forms are close to icons. They are often highlighted.
- **Adverbs of intensity**: adverbs of intensity (e.g. *very* or of affimation (e.g. *certainly*), when applied to action verbs also introduce levels of strength *we strongly advise you not to buy..., this will certainly break ....*

We also noted forms that weaken the argument. For example, the presence of a positively oriented support and a negatively oriented one for a given instruction shows the pros and cons without developing too strong a positive or negative orientation. This may be viewed also as a subtle form of persuation where a kind of objective analysis is provided to the reader.

The above linguistic marks are quite stable over a large set of types of procedural texts. Some are more frequent in some types of texts, for example, marks related to typography and text visualisation are more frequent on the web for large public audiences. Those marks can be combined to stress supports more strongly. However, we observed that, in most cases, a maximum of two of these categories may be used jointly: beyond this level supports loose their effect.

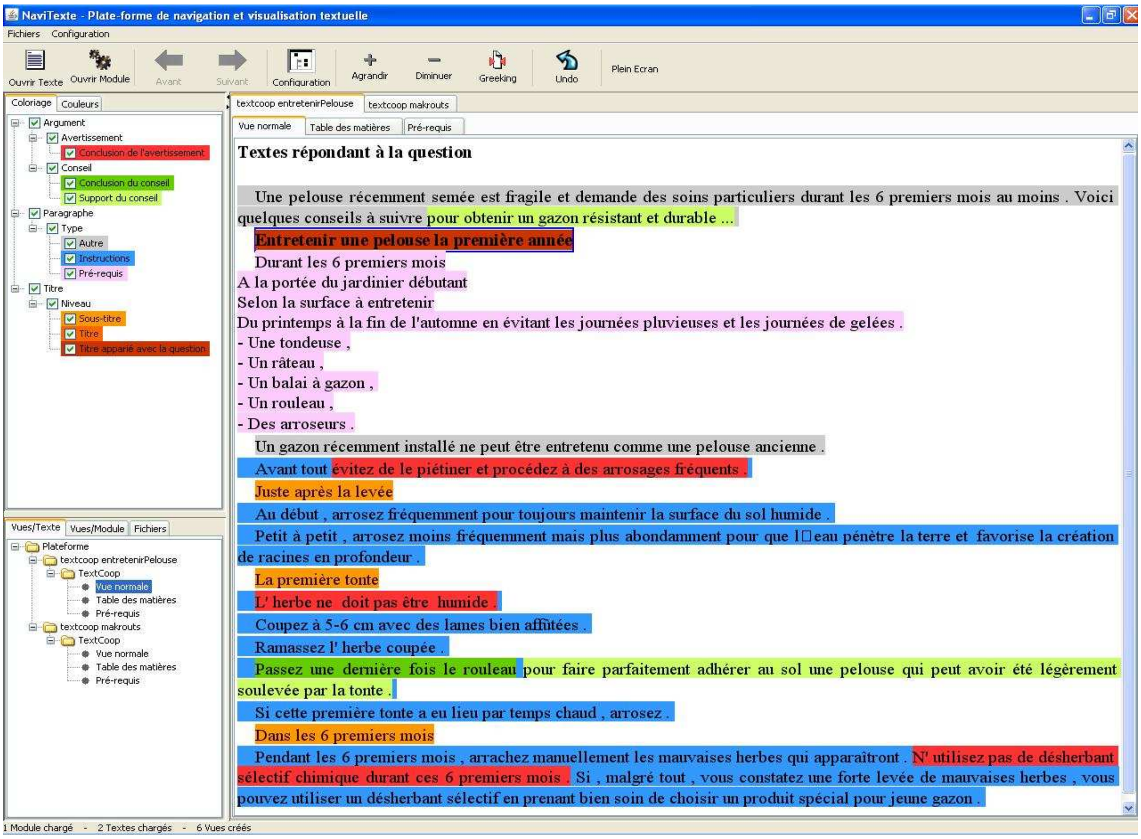
For each of these categories, we can tentatively define scales, but this is quite arbitrary and subject to errors. Research in lexical semantics, originating from (Cruse 1986) proposed some schemas for organizing along scales collections of terms which exhibit various levels of strength for a given property. However, we feel that, for each domain, these scales need to be constructed from complex and heavy psycho-linguistics experiments. We indeed noted that the relative importance of the strength of terms do depend quite heavily on the domain at stake and on the author of the text and the target audience. Obviously this is a task worth pursuing over some domains.

In a text where, in general, several arguments are found, the strength of an argument must also be evaluated w.r.t. the global strength of the others. This would be a useful contribution to Action Theory.

4.1 Perspectives

In this paper, we presented the different forms arguments and their associated persuasive forces may take in a large variety of procedural texts. We have developed several natural language patterns to recognize conclusions and supports and related persuation marks, with quite good an accuracy. Persuation marks cover a quite large spectrum of devices, from icons, punctuation, to more semantic aspects such as verb classes, and to pragmatic aspects.

This is obviously only a first step in the analysis process, since the heart of the problem is to be able to effectively measure the persuation force associated with an argument, in isolation and in relation with the other arguments in the procedure. At the moment, we can simply, based on patterns, say if the argument has a strong positive or negative orientation. We also gave a few syntactic and morphological factors that tend to reinforce this first evaluation.



Textes répondant à la question

Une pelouse récemment semée est fragile et demande des soins particuliers durant les 6 premiers mois au moins . Voici quelques conseils à suivre pour obtenir un gazon résistant et durable ...

Entretenir une pelouse la première année

Durant les 6 premiers mois
A la portée du jardinier débutant
Selon la surface à entretenir
Du printemps à la fin de l'automne en évitant les journées pluvieuses et les journées de gelées .
- Une tondeuse ,
- Un râteau ,
- Un balai à gazon ,
- Un rouleau ,
- Des arroseurs .

Un gazon récemment installé ne peut être entretenu comme une pelouse ancienne .

Avant tout évitez de le piétiner et procédez à des arrosages fréquents .

Juste après la levée

Au début , arrosez fréquemment pour toujours maintenir la surface du sol humide .

Petit à petit , arrosez moins fréquemment mais plus abondamment pour que l□eau pénètre la terre et favorise la création de racines en profondeur .

La première tonte

L' herbe ne doit pas être humide .

Coupez à 5-6 cm avec des lames bien affûtées .

Ramassez l' herbe coupée .

Passez une dernière fois le rouleau pour faire parfaitement adhérer au sol une pelouse qui peut avoir été légèrement soulevée par la tonte .

Si cette première tonte a eu lieu par temps chaud , arrosez .

Dans les 6 premiers mois

Pendant les 6 premiers mois , arrachez manuellement les mauvaises herbes qui apparaîtront . N' utilisez pas de désherbant sélectif chimique durant ces 6 premiers mois . Si , malgré tout , vous constatez une forte levée de mauvaises herbes , vous pouvez utiliser un désherbant sélectif en prenant bien soin de choisir un produit spécial pour jeune gazon .

Navitexte Output

References

Amgoud, L., Parsons, S., Maudet, N., *Arguments, Dialogue, and Negotiation,* in: 14th European Conference on Artificial Intelligence, Berlin, 2001.

Chesnevar, C., et alii., it Towards an Argument Interchange Format, The Knowledge Engineering Review, 2007, Cambridge University Press.

Cruse, A., lexical Semantics, Cambridge University Press, 1986.

Dautriche, I. Saint-Dizier, P., *A Conceptual and Operational Model for Procedural Texts and its Use in Textual Integration*, IWCS8, Tilburg, January 2009.

Delpech, E., Saint-Dizier, P., Investigating the Structure of Procedural Texts for Answering How-to Questions, LREC 2008, Marrakech.

Fontan, L., Saint-Dizier, P., Analyzing the explanation structure of procedural texts: dealing with advice and Warnings, STEP conference, Venice, August 2008.

Grosz, B., Sidner, C., Attention, intention and the structure of discourse, Computational Linguistics 12(3), 1986.

Kosseim, L., Lapalme, G., *Choosing Rhetorical Structures to Plan Instructional Texts,* Computational Intelligence, B. Blackwell, Boston, 2000.

Mann, W., Thompson, S., *Rhetorical Structure Theory: Towards a Functional Theory of Text Organisation*, TEXT 8 (3) pp. 243-281, 1988.

Moens, M-F , Boiy, E. , Mochales Palau R. , Reed, C., *Automatic Detection of Arguments in Legal Texts*, in Proceedings of the Eleventh International Conference on Artificial Intelligence and Law, ACM Press, NY, 2007.

Moschler, J., *Argumentation et Conversation*, Hatier - Crédif, 1985.

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

Van der Linden, K., *Speaking of Actions Choosing Rhetorical Status and Grammatical Form in Instructional Text Generation* Thesis, University of Colorado, 1993.

Walton, D., Reed, C., Macagno, F. (eds), *Argumentation Schemes,* Cambridge University Press, 2008.

Realm Traversal In Biological Discourse: From Model To Experiment and Back Again

Anita de Waard (1, 2)

(1) Elsevier Labs, Radarweg 29, 1043 NX Amsterdam, the Netherlands
a.dewaard@elsevier.com
(2) UiL-OTS, Utrecht University, Utrecht, the Netherlands

Abstract

We investigate the linguistic manifestations of scientific sensemaking in experimental research papers by performing a discourse analysis at the subsentential level. Our analysis consists of, first, segmenting the texts roughly at clause level, and next, identifying a set of segment types, describing the various rhetorical elements that make up a scientific text. To enable a unilateral definition of these segment types, we study correlations between these segment types and specific linguistic features, such as verb form (tense, aspect, and nonfinite forms), semantic verb class, the presence of first-person pronouns, and a number of other features. Here, we investigate a specific rhetorical action, namely the transition between conceptual and experimental elements within a research article and find four ways that this transition is encoded. We find four manifestations of these transitions: through verb form change, by use of two different segment types, and by means of sentence-initial phrases.

Keywords: Discourse analysis, verb tense, experimental research articles, discourse segmentation.

1 A two-dimensional model of biological discourse

There is a general agreement on the structure of the experimental research paper as it has evolved over the last 300 years (Bazerman, 1988; Latour, 1997; Biber, 2005; Swales, 1990, 2004; Hyland, 2004). Largely, there are four sections to such papers, that are so common they have been converted to an acronym: IMRaD, or Introduction, Methods, Results and Discussion. Overall, the paper starts with an Introduction, where the scene is set, a research space is 'created' (Swales, 1990, 2004), and the overall research question is posed. The next two sections, Methods and Results, are very different in character: they describe the experiments performed by the authors, the methods used, and the results found, often represented in the form of figures and references. In the Discussion section conclusions are drawn from the experiments pertaining to the research question discussed in the Introduction, and the experimental results are given their conceptual interpretation. Therefore, as a global outline, the paper moves from discussing concepts to discussing experiments, and concludes with a mostly conceptual focus. But if we look at a more fine-grained level, we see that experiments are also addressed in the Introduction, and theoretical concepts are described and mentioned in the Results, and, to a lesser degree in the Methods section. Thus, an experimental research article is always active on two levels: the conceptual and the experimental.

For an article to be successful, it is imperative that the reader be convinced that the experiments done pertain to the conceptual research questions, and their outcomes support the author's conclusions. It is also important, to satisfy genre characteristic modes of evidence creation, that the authors are always clear which statements pertain to their own experiments, and which to general theoretical concepts. Therefore, the transitions between these two realms, the experimental and the conceptual, form a critical element of the persuasive discourse. So how does the reader make sense of a text that simultaneously develops a narrative on these two, quite different, levels? And how is transfer of the reader's attention between these two levels achieved?

To investigate this issue, let us first look at an example; (1)-(4) are consecutive sentences taken from the Results section of a cell biology paper (Voorhoeve et al., 2006):

(1) Oncogene-induced senescence is characterized by the appearance of cells with a flat morphology that express senescence associated (SA)-β-Galactosidase.

(2) Indeed, control RAS^{V12}-arrested cells showed relatively high abundance of flat cells expressing SA-β-Galactosidase (Figures 2G and 2H).

(3) Consistent with the cell growth assay, very few cells showed senescent morphology when transduced with either miR-Vec-371&2 or control p53^{kd}.

(4) Altogether, these data show that transduction with either miR-Vec-371&2 or miR-Vec-373 prevents RAS^{V12}-induced growth arrest in primary human cells.

Here, (1) describes a theoretical or conceptual bit of knowledge, generally assumed to be known. Sentence (2) describes an experimental result. In (3) the experimental results are still discussed, but in (4) we transfer 'back' to a statement of a more general nature, that draws a conclusion about these experimental results. So we see a transfer from the conceptual realm to the experimental, between (1) and (2), and back again, between (3) and (4). The question we will investigate in this paper is: how are these transitions linguistically marked? To study this issue, we have identified a syntactically motivated segmentation of the text into discourse units semantic segment types (discussed in Section 2). We describe 4 mechanisms by which this occurs in Section 3; in Section 4 we discuss our results.

2 Discourse Segments

2.1 Segmentation

To come to an understanding of how the experimental and conceptual realms are defined and transitioned, within a text, and even within a sentence, we need to develop a way of identifying the rhetorical moves that occur, by identifying segments of text that perform a distinct pragmatic or rhetorical function (see e.g. Swales, 1990, 2004; Biber, 2005). There are many ways to chunk a text into non-overlapping segments (see e.g. Carlson, et. al, 2001; and Pander Maat, 2002). To allow a reproducible method of segmentation that can be syntactically motivated, we have developed a set of grammatical criteria. For a full description, see (de Waard and Pander Maat, Preliminary Report): as a summary, we define as a segment:

- Simple sentences;
- Coordinated clauses;
- Matrix clauses;
- Clauses that function as the Direct Object of a sentence, and contain a Subject;
- Clauses that function in an Adverbial role;
- Sentential Relative clauses;
- Finite, Non-restrictive Noun-Phrase postmodifiers.

In contrast, Verbless clauses and Clauses that function as the Subject of a sentence are not segmented.

2.2 Segment Types

After dividing the text up into these segments, we have empirically defined a set of semantic segment types, described in more detail in (de Waard and Kircz, 2007). A summary of basic segment types is:

- **Fact**: a claim that has been accepted to be true, a known fact.
- **Hypothesis**: a proposed idea, not supported by evidence.
- **Problem**: unresolved, contradictory, or unclear issue.
- **Method**: experimental method
- **Result**: a restatement of the outcome of an experiment
- **Implication**: an interpretation of the results, in light of earlier hypotheses and facts
- **Goal**: research goal

Of these, the first three refer to the *conceptual realm*: known facts, problems and hypotheses are used to describe the current model and plans for new models. Method and Result clearly describe the *experimental findings*. The last two segment types are more ambiguous. Implication is a special type of segment, that discusses the conclusions drawn from the experiments – it is a conceptual segment itself, but is usually directly preceded by a Results segment (see de Waard and Kircz, 2007, for a list of segment orders). The Goal segment usually states the opposite transition: in stating the way that the conceptual hypotheses are tested, it offers a transition into the experimental realm (see Section 3.3 for more details on this role).

Next to these basic segment types, we define two derived classes of segments. The first division is between ‘own’ and ‘other’ segments, i.e. bits of text discussing the authors’ own Results/Goals/Methods/Implications, and those of others. For example, ‘Drosophila germ cells and mouse embryonic stem (ES) cells require miRNAs to proliferate (Forstemann et al, 2005 and Hatfield et al, 2005).' is an 'Other-Implication' segment. In terms of the Experiment/Concept division, Other-segments 'inherit' the realm they discuss; Other-Hypotheses and Other-Implications still concern the conceptual realm, just as Methods and Results concern Experiments.

The second derived set of segments are metatextual or Regulatory (Degand, 2005) segments. These bits of 'text about text' serve to regulate 'core' statements, by either drawing the reader's attention, or by offering an authorial stance about the preceding and introducing the succeeding segment. An example is the segment 'These results suggest that....', which we denote a 'Regulatory-Implication' segment (since it is followed by an Implication). These segments are often (but not always used) to indicate a realm transition; therefore, we will denote them 'transitional' segments.

Two other segment types have to do with the text structuring itself: intratextual segments (e.g., 'see Figure 4') and intertextual segments (e.g., 'Reviewed in Agami et al.') There is no direct connection to either realm; therefore, these will be denoted as pertaining to the 'textual' realm. When we apply this segmentation, sentences (1)-(4) look as follows (see Table 1).

Nr.	Segment	Segment Type	Realm
(1) a.	Oncogene-induced senescence is characterized by the appearance of cells with a flat morphology	Fact	Concept
(1) b.	that express senescence associated (SA)-β-Galactosidase.	Fact	Concept
(2)	Indeed, control RASV12-arrested cells showed relatively high abundance of flat cells expressing SA-β-Galactosidase (Figs 2G-H).	Result	Experiment
(3) a.	Consistent with the cell growth assay, very few cells showed senescent morphology	Result	Experiment
(3) b.	when transduced with either miR-Vec-371&2, miR-Vec-373, or control p53kd.	Method	Experiment
(4) a.	Altogether, these data show that	Reg-Implication	Transfer Exp-> Concept
(4) b.	transduction with either miR-Vec-371&2 or miR-Vec-373 prevents RASV12-induced growth arrest in primary human cells.	Implication	Concept

Table 1: Segmentation and segment types for (1) – (4).

The transfer between the conceptual realm and the experimental occurs between (1) and (2), and between (3) and (4), via (4)a. We will now discuss various mechanisms for this 'realm transfer'.

3 Realm Transitions

For our study, we have analysed two full-text biology articles, (Voorhoeve et al., 2006) and (Louiseau et al, 2008), and identified in which sections which segments occur in (Table 2). We admit that this is a very limited data set, and our efforts are to expand this; however, some trends can be made visible even when looking at only two texts. We will now discuss are in a position to discuss the typical differences, and the transitions, between segments that have a conceptual, and an experimental focus. To do so, we have studied several transitions between these segment types. We have identified four mechanisms by which these transitions occur, which we now discuss in turn.

3.1 Verb form change.

One of the features we have studied for each segment type is that of verb form, for finite and nonfinite forms. Elsewhere, we motivate the verb forms we study (De Waard and Pander Maat, Preliminary Report); a summary of the results per segment type is given in Table 3.

	Present	*%*	**Pres Perfect**	*%*	**Past**	*%*	**Future**	**To-Infinitive**	**Gerund**	**Past Part**	***Total Other***	*%*	**Total**
Fact	41	*91%*	0	*0%*	0	*0%*	1	0	1	2	*4*	*9%*	**45**
Problem	23	*74%*	2	*6%*	3	*10%*	0	2	1	0	*3*	*10%*	**31**
Hypothesis	25	*71%*	0	*0%*	1	*14%*	0	3	2	0	*5*	*14%*	**34**
Other-Hypothesis	4	*50%*	1	*13%*	2	*13%*	0	2	0	0	*2*	*25%*	**9**
Implication	55	*63%*	1	*1%*	9	*10%*	0	6	16	1	*23*	*26%*	**88**
Other-Implication	17	*61%*	5	*18%*	5	*18%*	0	1	0	0	*1*	*4%*	**28**
Total Concept	**165**	***70%***	**9**	***4%***	**23**	***10%***	**1**	**14**	**20**	**3**	***38***	***16%***	**235**
Method	6	*4%*	0	*0%*	109	*73%*	0	3	28	3	*34*	*23%*	**149**
Other-Method	1	*25%*	0	*0%*	1	*25%*	0	0	1	1	*2*	*50%*	**4**
Result	12	*9%*	0	*0%*	112	*85%*	0	6	0	1	*7*	*5%*	**131**
Other-Result	27	*53%*	2	*4%*	19	*37%*	2	0	1	0	*3*	*6%*	**51**
Total Experiment	**46**	***14%***	**2**	***1%***	**241**	*72%*	**2**	**9**	**30**	**5**	***46***	***14%***	**335**
Goal	3	*8%*	2	*5%*	8	*20%*	0	27	0	0	*27*	*68%*	**40**
Reg-Hypothesis	8	*67%*	0	*0%*	3	*25%*	0	0	1	0	*1*	*8%*	**12**
Reg-Problem	3	*75%*	0	*0%*	0	*0%*	0	0	1	0	*1*	*25%*	**4**
Reg-Implication	31	*78%*	2	*5%*	1	*3%*	0	0	6	0	*6*	*15%*	**40**
Reg-Result	1	*7%*	0	*0%*	9	*64%*	0	0	1	3	*4*	*29%*	**14**
Total Transition	**46**	***42%***	**4**	***4%***	**21**	***19%***	**0**	**27**	**9**	**3**	***39***	***35%***	**110**
Intertextual	3	*25%*	2	*17%*	2	*17%*	0	0	0	5	*5*	*42%*	**12**
Intratextual	6	*60%*	0	*0%*	4	*40%*	0	0	0	0	*0*	*0%*	**10**
Total Text	**9**	***41%***	**2**	***9%***	**6**	***27%***	**0**	**0**	**0**	**5**	***5***	*23%*	**22**
Total	***266***	***38%***	***17***	***2%***	***291***	***41%***	***3***	***50***	***59***	***16***	***128***	***18%***	***702***

Table 3: Verb form vs. segment category for Louiseau and Voorhoeve

We see that indeed, in general, there is a strong correlation between verb form seems and whether a segment concerns the conceptual realm, referred to in the present, or the experimental realm, referred to in the past.

There are some interesting correlations: for instance, ‘Other-’ segments are often stated in the Present Perfect. Regulatory segments occur both in the past and present tense; a possible problem can be that we have not counted ‘Reg-Other-clauses’, and therefore there might be clauses where past work was discussed, that might lead to e.g. a high incidence in the use of the Past tense for Regulatory-Hypothesis segments. But we really need to have more data to come to any serious conclusions on these points.

When we look at the non-finite clauses, there is a clear correlation for a “goal’ segment to be stated as a to-infinitive; again, since these clauses are also referred to as purpose clauses

(Quirk et al., s. 15.48), this is not surprising. Interestingly enough, Methods and Implications segments show a high incidence of being stated in the gerund. For the Method, this is understandable enough; in a phrase such as 'following RASV12 induction'. The gerund in the Implication segments usually (in 12 of 16 cases) is for the verbs 'revealing', 'resulting' or 'indicating'. Therefore, the gerund with this specific type of verb often signals a transition to an Implication segment, in a clause like '*revealing* a lack of improved recognition...'

So is verb form change a clear marker for realm change? The correlation between verb form and realm seems to indicate that this would be the case. When we look at some examples, we do indeed see a changeover from experimental description from concepts by change in verb tense, such as between sentences (1) and (2).

In some cases, this transfer can even occur several times within a sentence:

(5) a. Only cells **infected** with the pSM1 library

(5) b. **exhibited** formation of macroscopic colonies in semisolid media (Figure 1D, right panels),

(5) c. **indicating** the presence of shRNAs

(5) d. that **transform** TLM-HMECs.

We see that (5) a. and b., which respectively refer to the experimental Method and Result, are given as a past participle, and in the past tense; (5) c. is again a 'Regulatory-Implication' segment, using a gerund form of a causal verb, as discussed above; and (5) d., a Noun-Phrase postmodifier to 'shRNAs', is a Fact segment; this does concern the conceptual realm (as it describes a known fact about the shRNAs), and is stated in the present tense. So, we see two verb form changes, between (5) b. and (5) c., which indicate the start of a realm change; the change from (5) b. to (5) d. is indeed a 'complete' realm change.

To adhere a small quantification for this claim, we counted the total number of realm changes vs. the number of verb form changes in a single paper (Voorhoeve 2006), and arrive at Table 4:

	Verb form change	**No Verb form change**
Realm Change	97	3
No Realm Change	20	(not relevant)

Table 4: Verb form change vs. number of experimental-conceptual realm transitions in Voorhoeve (2006)

From Table 4, we see that in the overwhelming majority of the cases (97 % of the realm transitions (which are accidentally exactly 100!)), a verb form change accompanies a realm change. In 9 + 21 = 30% of the cases, there is a specific segment type associated with the transition. Below, we discuss some examples with Goal and Regulatory transitions.

There are, of course, anomalies. A different example is given in (6):

(6) a. Figure 1C shows that

(6) b. cells transduced with miR-Vec-24 clearly **express** high levels of mature miR-24,

(6) c. whereas little expression **was detected** in control-transduced cells.

where (6) b. expresses a Result, in the present tense. This might be a result of the fact that Dutch native speakers tend to have a higher incidence of present tense than the average (Burrough-Boenisch, 2003) and could also indicate a rhetorical highlighting of this claim.

3.2 Goal /Methods segments

A further way in which to indicate the transfer from the conceptual to the experimental realm is by using a Goal or Methods segment. As an example, the transfer to the experimental realm in (8) is achieved by a typical to-infinitive Goal segment:

(7) This escape from oncogene-induced senescence is a prerequisite for full transformation into tumor cells.
(8) a. To identify miRNAs that can interfere with this process
(8) b. and thus might contribute to the development of tumor cells,
(8) c. we tran[s]duced BJ/ET fibroblasts with miR-Lib
(8) d. and subsequently transduced them with either RASV12 or a control vector (Fig. 2B).

A different example is provided by a Methods segment, e.g. (10) a. is a methods segment that transfers back from the conceptual to the experimental realm.

(9) The cancerous process can be modeled by in vitro neoplastic transformation assays in primary human cells (Hahn et al, 1999).
(10) a. Using this system,
(10) b. sets of genetic elements required for transformation were identified.

We see two verb form changes: the typical gerund for the Methods section, and then a transfer to the experimental past tense. There is an analogy with Goals: here, the typical verb form is that of the to-infinitive. It would be interesting to study other occurrences of these nonfinite verb forms as transition to different textual realms.

3.3 Regulatory segments

To move in the other direction, - from the experimental to the conceptual realm – a very common construction is to use a segment such as (4)a, of the type 'these data show that', which we have typed Regulatory sentences, usually Regulatory-Implication or Regulatory-Hypothesis (rarely, Reg-Problem). Their typical structure is "these/this {result/observation(s)) show/imply/suggest/indicate/demonstrate (s) that,", with a very small list of verbs used: indicate (9x), suggest (5x), show (3x), demonstrate (3x), to be significant, to point to, to validate (each 1x). Generally, these segments are stated in the present tense, but occasionally using the verbs indicate or to validate, are given as an –ing participle, 'validating that'. These constructions are very common, occurring 30 times in the Voorhoeve paper) and always indicate a transition from the experimental to the conceptual realm.

In table 5, we have indicated the number of times that a tense change is accompanied by the segment types discussed above:

	Verb form change	**No Verb form change**
Verb form only	67	
Goal segment	9	
Regulatory segment	21	3
Total Realm Change	**97**	**3**

Fig 5. Verb form change vs. segment-indicated change for experimental-conceptual realm transitions in Voorhoeve (2006)

3.4 Sentence-initial adverbials

An often-encountered method of signaling text transitions is through the use of sentence-initial adverbials (Ho-Dac, Lydia-Mai & Péry-Woodley, 2008). Do these play a role in signaling realm transitions in biology text, as well? We have not performed an in-depth study of this issue, but a first analysis of the Voorhoeve paper seems to indicate that they sometimes do, but sometimes don't. For example, (2) and (4), which both signal a realm transfer, also contain a sentence-initial adverbial ('Indeed' and 'Altogether'). Often, sentence-initial adverbials are used together with a verb form-change (as in (2) or with a regulatory segment (as in (4) a). In Voorhoeve et al. (2006), the following sentence-initial adverbials indicated a realm change:

therefore (6x), altogether (3x), indeed (3x), interestingly (3x), next (3x), furthermore, here, importantly, moreover, to this end;

and the following did not indicate a realm change:

whereas(10x), however (6x), indeed (5x), in contrast (4x), furthermore (3x), significantly (3x), therefore (3x), while (3x), additionally, finally, first, as expected, in accordance, in general, moreover, nevertheless, interestingly, previously, since, with one exception.

There is not a clear distinction between the Sentence Initial Adverbials that do or do not signal a realm change. An explanation could be that elaborations (furthermore, additionally) perhaps are less likely to indicate a realm change, since they indicate a continuation of an argument or discussion, whereas causal connectives (therefore, altogether) are perhaps more likely to do indicate realm changes, since they identify a relationship between a claim (usually conceptual) and its evidence (often experimental). This merits further study.

4 Discussion

In conclusion, discourse segmentation and segmentation typing seem to be a useful tool to study the transitions between various discourse realms in scientific text, and we postulate that a change in verb form is an important way to obtain this transfer. If we consider the mental space metaphor, we propose that within a scientific paper, a 'model space' is opened in the Introduction of a paper, which is kept open throughout the course of the paper. After opening this background space, the experimental experiences of the authors are narrated as a series of events that occurred in the past. Whenever reference is made to the theoretical realm, present tense re-invoked the space created there; conversely, in a later section of the paper (e.g. the Discussion), the experiments are referred back to by using the past tense. We do not have enough data to draw convincing conclusions, and more research is merited on this issue. We are currently engaged in a series of preliminary investigations to further quantify and apply these results, to see if finding the concept-experiment boundary can help access the core argumentation in a biology text.

Furthermore, we have only discussed transitions between the experimental and conceptual realm here: spoken in discourse processing terms, we have only addressed the 'situational model'; what the text is *about*. However, some of the verb form changes might indicate other realms that are addressed and activated during the reading of a text. For instance, in (12), the intertextual segment 'Note that' indicates a jump from the topic under description, to a direct address of the reader:

(11) a. As demonstrated by RPA,

(11) b. miR-Vec-372mut and miR-Vec-373mut indeed failed to express miR-372 and miR-373, respectively (Figure 3A).

(12) a. Note that

(12) b. miR-Vec-372-mut still expressed miR-371 to a similar extent as the original miR-Vec-371&2.

Briefly put, the narrative of the text itself could be seen as separate realm, that is possibly also addressed, 'opened', using verb forms. We hope to elucidate a more detailed model taking into account more dimensions of scientific discourse in our future work.

References

BAZERMAN, C. (1988). *Shaping written knowledge: the genre and activity of the experimental article in science*. Madison, Wisconsin: Univ. of Wisconsin Press, 1988.

BIBER, D., & JONES, J. K. (2005). Merging corpus linguistic and discourse analytic research goals: Discourse units in biology research articles. *Corpus Linguistics and Linguistic Theory*, 2(2005).

BURROUGH-BOENISCH, J. (2003). Examining present tense conventions in scientific writing in the light of reader reactions to three Dutch-authored discussions. *English for Specific Purposes*, Volume 22, Issue 1, 2003, Pages 5-24.

CARLSON, L., MARCU, D., AMORRORTU, E., HOBBS, J., KOVARIK, J., MERRIKEN, T., ET AL. (2001). Discourse Tagging Reference Manual. *Structure*, (2), 1-87.

DEGAND, L., & CATHERINE, A. (2005). Minimal Discourse Units: Can we define them, and why should we? In: Aurnague, M., Bras, M., Le Draoulec, A., & Vieu, L. (eds). Proceedings of *SEM-05. Connectors, discourse framing and discourse structure: from corpus-based and experimental analyses to discourse theories*, Biarritz, 14-15 November 2005, 65-74.

DE WAARD, A. AND KIRCZ, J.G. (2008). Modeling scientific discourse - shifting perspectives and persistent issues, ELPUB2008. Open Scholarship: Authority, Community, and Sustainability in the Age of Web 2.0 – Proc. of the 12th Int. Conference on *Electronic Publishing*, June 2008, Eds. L. Chan and S. Mornati, pp. 234-245

DE WAARD, A. & PANDER MAAT, H. (2009). Discourse segmentation of biology texts, Preliminary Report, UiL-OTS, http://elsatglabs.com/labs/anita/papers/Segments.pdf

HO-DAC, L-M& PÉRY-WOODLEY, M-P. (2008). Temporal adverbials and discourse segmentation revisited. 2008. W. Ramm & C. Fabricius-Hansen (eds.)*'Linearisation and Segmentation in Discourse. Multidisciplinary Approaches to Discourse 2008 (MAD 08),'* Feb 20-23 2008, Lysebu, Oslo. Oslo: Dept. of Literature, Area Studies and Europ. Languages, Univ. of Oslo

HYLAND, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing*. Addison Wesley Publishing Company, 2004.

LATOUR, B., AND WOOLGAR, S. (1979). *Laboratory Life: The Social Construction of Scientific Facts*. Beverly Hills: Sage, 1979.

LOISEAU, F., MILLAN, M.J. (2009). Blockade of dopamine D3 receptors in frontal cortex, but not in sub-cortical structures, enhances social recognition in rats: Similar actions of D1 receptor agonists, but not of D2 antagonists. *European Neuropsychopharmacology* - January 2009 (Vol. 19, Issue 1, Pages 23-33).

PANDER MAAT, H. (2002). *Tekstanalyse. Wat teksten tot teksten maakt.* Bussum: Coutinho 2002.

QUIRK, R., GREENBAUM, S. LEECH, G. , SVARTVIK, J. (1985) *A comprehensive grammar of the English language.* Addison and Wesley, 1985.

SWALES, J.M. (1990). *Genre Analysis*, Cambridge: Cambridge University Press.

SWALES, J.M. (2004). *Research genres: explorations and applications*. Cambridge University Press, 2004.

VOORHOEVE PM, LE SAGE C, SCHRIER M, GILLIS AJ, STOOP H, NAGEL R, LIU YP, VAN DUIJSE J, DROST J, GRIEKSPOOR A, ZLOTORYNSKI E, YABUTA N, DE VITA G, NOJIMA H, LOOIJENGA LH, AGAMI R. (2006). A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. Cell. 2006 Mar 24;124(6):1169-81.

Effects on Text Processing of Signaling Text Organization

Robert F. Lorch, Jr.
Department of Psychology, University of Kentucky
rlorch@email.uky.edu

Keywords. Signaling, Organizational Signaling; Text-Processing

1 Defining the domain

The talks at the 2010 Conference on Multidisciplinary Approaches to Discourse address a wide range of senses of the term "signaling." My use of "signaling" will be a bit more restrictive than the sense in which many researchers use the term. Loosely, I consider a signal to be a writing device that emphasizes aspects of a text's organization and/or content, but which can be omitted from the text without compromising the cohesiveness of the text (Lorch, 1989). This definition contrasts with Meyer's (1975) definition of text signals, which is tied to her system of classifying prototypical expository text structures. In particular, notable omissions from our definition of signals are connectives and what Meyer calls "keywords." Nevertheless, our definition still admits a diverse collection of writing devices, including:

- Overviews, outlines, and summaries
- Systems of headings and individual headings
- Bulleting, including enumeration
- Segmentation cues
- Preview and summary sentences
- Function indicators and importance indicators (e.g., emphasis)
- Cross-reference

There are, of course, alternative approaches to the study of signaling devices, many of which are well-represented at this conference. A linguistic approach naturally focuses on questions about the linguistic functions served by various signals (e.g., Behrens &amp; Solfjeld; Roze, Danlos &amp; Muller). Also within this approach are questions about how writers and speakers actually use signals (e.g., Longrée &amp; Mellet; Martin, Delin &amp; Waller). Educational psychologists have long focused on the question of how signals might be used to improve learning (e.g., Fabre, Ho Dac, Péry-Woodley, &amp; Rebeyrolle). Researchers concerned with text design and formatting might ask how to best match signals to the purpose of the text (e.g., Zafiharimalala &amp; Tricot; Waller &amp; Delin). And cognitive scientists focus on questions about how signals influence the processing of text by the reader (e.g., Giraud &amp;

Thérouanne; Laippala; Canestrelli, Sanders, Mak). As signaled in the time of my talk, this general question has been the primary focus of my research. My goals are (1) to share with you my perspective on the signaling research that has been conducted by psychologists and (2) to present a theoretical framework that I have developed with my colleagues at the Université de Toulouse.

2 Some psychological research on signaling outcomes

Empirical research on signaling in psychology goes back to the 1950s (Christensen &amp; Stordhal, 1955; Klare, Mabry &amp;amp; Gustafson, 1955; Klare, Shuford &amp; Nichols, 1958). Most of the research has been conducted by researchers with interests in education, so the major goal of the research has been to identify effects of signaling that might be used to improve instruction. Most of the research has assessed signaling effects using memory measures, particularly free recall of text. For example, a typical experiment might compare recall of a text containing headings with recall of the same text without the headings.

Early research on the effects of organizational signals (e.g., headings, overviews) on memory for text was motivated, in part, by findings from list-learning experiments that organizational factors have large effects on memory for lists of words. In this context, it was surprising to find that many of the early studies of signaling found little influence of organizational signals on memory for text. Subsequent research suggested two reasons for the early null effects. One is that the texts used in early research were probably too short and too simply structured to reveal signaling effects. The other is that signaling effects on text recall are more complex than early researchers anticipated.

There is good evidence that organizational signals benefit memory for text only when readers' background knowledge and text-processing strategies require additional support to meet the challenges presented by a text. There are several examples. Novices in a domain depend heavily on signaling contained in a science text whereas students with sufficient background are not influenced by the same signaling (Dee-Lucas &amp; Larkin, 1986; 1988). Organizational signals benefit recall of content associated with unfamiliar text topics, but not recall of familiar topics (Lorch &amp; Lorch, 1996a); they benefit recall of relatively unelaborated topics more than elaborated topics (Lorch, Lorch &amp; Inman, 1993); they benefit recall of poorly organized texts more than well-organized texts (Lorch &amp; Lorch, 1985); and they benefit recall of complexly-organized texts more than simply organized texts (Lorch, Lorch &amp; Matthews, 1985). In short, the more challenges a text presents to a reader's ability to understand, encode, and recall the text's organization, the more beneficial signaling is.

When signals do benefit memory, the benefits are not always in terms of better overall recall. Rather, organizational signals consistently influence the distribution of recall of text content. Specifically, when organizational signals are omitted from a text, readers tend to recall the initial content of the text well and perhaps some of the concluding content, but they forget entire sections of the text. In contrast, when organizational signals are added to the same text, reader do a better job of remembering the text topics and their organization so their recalls are more complete in the sense that they include something about more of the topics in the texts (Loman &amp; Mayer, 1983; Lorch, 1996a; 1996b; Lorch et al., 1993; Mayer, Dyck &amp; Cook, 1984; Meyer, Brandt &amp; Bluth, 1984). These findings have been interpreted as indicating that organizational signals induce readers to use a different text processing strategy than they would otherwise use. Namely, the signals cause more attention to text topics and their organization with the result that readers created a mental outline – what I call a "topic structure representation" (TSR) – that is then available to guide memory retrieval at the time of recall.

3 Shifting the focus to how signals influence text processing

Much of the early psychological research on signaling was more focused on the question of what effects signals have than on the question of how signals achieve their effects. Research over the past 25 years has increasingly looked at the cognitive mechanisms underlying signaling effects. This change in focus has led the use of new ways of assessing signaling effects in addition to the conventional memory measures.

One way that we have attempted to look at influences of signals on readers' understanding of text structure has been to use a summarization task in which the text is available during summarization to reduce dependence on memory. These studies (Lorch &amp; Lorch, 1996a; Lorch, Lorch, Ritchey, McGovern &amp; Coleman, 2001) demonstrate that readers rely heavily on signals to identify topics and their organization; in the absence of organizational signals, they use secondary indicators (e.g., relative elaboration of topic; familiarity of topic). Further evidence for differences in how readers encode signaled vs. unsignaled text comes from the finding that when topic labels are used to cue recall, the cues lead to big improvements in recall (compared to free recall) when the text contains signals than when the text omits signals (Lorch &amp; Lorch, 1995).

We have attempted to study the influences of signals on readers' online processing of text. We have done this both using a sentence-by-sentence presentation of texts under the control of the reader, and by tracking eye movements during full page displays. One consistent finding from the sentence-by-sentence task is that signaling content with typographical contrast and signaling organization by enumeration leads to more careful attention to the signaled content (Lorch &amp; Chen, 1986; Lorch &amp; Lorch, 1986; Lorch, Lorch &amp; Klusewitz, 1995). When we have used eye-tracking to study the influence of headings on text-processing (Hyöna &amp; Lorch, 2004; see also: Cauchard, Eyrolle, Cellier &amp; Hyönä, *in press*), we have found evidence that readers use headings to identify new topics and to guide their processing of text structure (as revealed by patterns of regressions from the ends of text sections back to the sentence-initiating headings).

Finally, we have tested our hypothesis that organizational signals affect readers' encoding of text and their subsequent recall strategy by training readers to use the hypothesized encoding/retrieval strategy. We found that recall was indistinguishable for readers who were trained in the strategy or received a text with signals or both, and all three groups exhibited better recall and a different distribution of recall of content relative to a group that read an unsignaled text and did not receive strategy training (Sanchez, Lorch &amp; Lorch, 2001).

4 A theoretical framework for psychological research on signaling

Although our research program has yielded some insights into how organizational signals influence various aspects of text-processing, our progress has been hindered by the lack of a comprehensive theoretical framework. Consequently, we recently developed a framework that is now in the initial stages of empirical testing. The theory is called "SARA" for "**S**ignaling

Available **R**elevant **A**ccessible" information (Lemarié, Lorch, Eyrolle &amp; Virbel, 2008). The theory consists of both a linguistic analysis of signaling devices and a cognitive analysis of how signals are likely to influence various cognitive processes involved in text-processing. Central to the theory is its analysis of seven distinct types of information that signaling devices communicate:

- Demarcation of structural boundaries
- Identification of text topics
- Identification of function
- Labeling of text sections
- Communication of hierarchical organization
- Communication of sequential organization
- Emphasis

It is hypothesized that any signaling device is capable of serving one or more of these seven "information functions," and that an analysis of the information functions of signals captures important similarities and differences among signaling devices. Further, it is hypothesized that each of these distinct types of information that signals make *available* to a reader have different implications for cognitive processing. Thus, the information functions of signals are the starting point for understanding how signals affect cognitive processing.

The particular information that is made available by a signal in a given text remains the same regardless of who is reading the text or why they are reading it. However, not all readers have the same knowledge and abilities, and the goals of readers are not the same in all situations. Thus, a second critical consideration in understanding the cognitive effects of signals is the *relevance* of the signal to the reader. If a signal is to influence text processing, it must pertain to the reader's goals and it must provide information that supports the reader's processing of the text.

Finally, a signal may make available information that is relevant to the reader, yet the reader may not attend consistently to the signaled information or may find it difficult to utilize the signaled information. Thus, we make a distinction between the availability vs. accessibility of signaled information. Whereas availability refers to information provided by a signal, accessibility refers to the ease with which readers can use that information. In some cases, information that is made available by a signal may not be used because the cognitive demands of processing the information are prohibitive. In other cases, the processing strategy of the reader may cause relevant information to go unprocessed. In either situation, the signal will not influence text-processing despite the fact that it makes relevant information available to readers.

In sum, SARA asserts that signaling effects on cognitive processing are a joint function of the relevance and accessibility of the information a signal makes available. This theoretical framework calls for a shift in the type of questions asked about signaling. Until now, research on signaling has focused on signaling devices. That is, researchers have asked questions such as "how do headings influence memory for text?" and "how do outlines affect text comprehension?" SARA urges a change in focus from the signaling devices to the information functions that the signaling devices perform.

References

Cauchard, F., Eyrolle, H., Cellier, J.-M. & Hyönä, J. in press Vertical perceptual span and the processing of visual signals in reading. *International Journal of Psychology*

Christensen, M. & Stordhal, K. E. 1955 The effect of organizational aids on comprehension and retention. *Journal of Educational Psychology*, 46:65-74.

Dee-Lucas, D., & Larkin, J. H. 1986 Novice strategies for processing scientific texts. *Discourse Processes* , 9:329-354.

Dee-Lucas, D. & Larkin, J. H. 1988 Novice rules for assessing importance in science texts. *Journal of Memory and Language* , 27:288-308.

Hyöna, J. & Lorch, R. F. 2004 Effects of topic headings on text processing: Evidence from adult readers' eye fixation patterns. *Learning and Instruction* , 14:131-152.

Klare, G. R., Mabry, J. E. & Gustafson, L. M. 1955 The relationship of patterning (underlining) to immediate retention and to acceptability of technical material. *Journal of Applied Psychology* , 39:40-42.

Klare, G. R., Shuford, E. H. & Nichols, W. H. 1958 The relation of format organization to learning. *Educational Research Bulletin* , 37:39-45.

Lemarié, J., Lorch, R.F., Jr., Eyrolle, H. & Virbel, J. 2008 A text-based and reader-based theory of signaling. *Educational Psychologist* , 43;27-48.

Loman, N. L. & Mayer, R. E. 1983 Signaling techniques that increase the understandability of expository prose. *Journal of Educational Psychology* , 75, 402-412.

Lorch, R. F. 1989 Text signaling devices and their effects on reading and memory processes. *Educational Psychology Review* , 1:209-234.

Lorch, R. F. & Chen, A. H. 1986 Effects of number signals on reading and recall. *Journal of Educational Psychology* , 78:263-270.

Lorch, R. F. & Lorch, E. P. 1985 Topic structure representation and text recall. *Journal of Educational Psychology* , 77:137-148.

Lorch, R. F. & Lorch, E. P. 1986 On-line processing of summary and importance signals in reading. *Discourse Processes* , 9:489-496.

Lorch, R. F. & Lorch, E. P. 1995 Effects of organizational signals on text-processing strategies. *Journal of Educational Psychology* , 87:537-544.

Lorch, R. F. & Lorch, E. P. 1996a Effects of headings on text recall and summarization. *Contemporary Educational Psychology* , 21:261-278.

Lorch, R. F. & Lorch, E. P. 1996b Effects of organizational signals on free recall of expository texts. *Journal of Educational Psychology* , 88:38-48.

Lorch, R. F., Lorch, E. P. & Inman, W. E. 1993 Effects of signaling topic structure on text recall. *Journal of Educational Psychology* , 85:281-290.

Lorch, R. F., Lorch, E. P. & Klusewitz, M. A. 1995 Effects of typographical cues on reading and recall of text. *Contemporary Educational Psychology* , 20:51-64.

Lorch, R. F., Lorch, E. P. & Matthews, P. D. 1985 On-line processing of the topic structure of a text. *Journal of Memory and Language* , 24:350-362.

Lorch, R. F., Lorch, E. P., Ritchey, K., McGovern, L. & Coleman, D. 2001 Effects of headings on text summarization. *Contemporary Educational Psychology* , 26:171-191.

Mayer, R. E., Dyck, J. & Cook. L. K. 1984 Techniques that help readers build mental models from scientific text: Definitions pretraining and signaling. *Journal of Educational Psychology* , 76:1089-1105.

Meyer, B. J. F. 1975 *The organization of prose and its effecs on memory* , Amsterdam: North-Holland.

Meyer, B. J. F., Brandt, D. M. & Bluth, G. J. 1980 Use of top-level structure in text: Key for reading comprehension of ninth grade students. *Reading Research Quarterly* , 16:72-103.

Sanchez, R. P., Lorch, E. P. & Lorch, R. F. 2001 Effects of headings on text processing strategies. *Contemporary Educational Psychology* , 26:418-428.

0… Second… Finally… Marking and unmarking of items in sequential text organisation

Veronika Laippala

University of Turku, Finland
veronika.laippala@utu.fi

Abstract This article presents an on-going work on text sequences in French. These are structures composed of at least partially ordered items introduced by markers of addition or order (*First… Then… Finally…/ D'abord… Ensuite… Enfin…).* The article concentrates on a subtype of text sequences, where at least one of the items is unmarked, i.e. not introduced by a marker of addition or order. The article aims at comparing these partially unmarked sequences to sequences with marked items only and to examine the unmarking more in detail.

The study shows that partially unmarked text sequences are often introduced by vague markers (e.g. *moreover / de plus*), whereas sequences with marked items only are frequently signalled by exact markers indicating order (*second / deuxièmement)*. Partially unmarked sequences are also longer than the other structures. The marking of text organisation appears therefore to be less explicit and vaguer in longer text segments.

Keywords: text sequences, enumerations, xml annotation, text organisation, cohesion, coherence.

1 Introduction

This article presents an on-going work on the marking of text organisation in research articles in French. The study concentrates on a particular type of organisation that is realised by text sequences. These are structures composed of at least partially ordered items of which at least some are introduced by markers of addition or enumeration: *The first example… 0… Finally… / Le premier exemple… 0… Enfin…* The corpus of the study consists of 90 research articles in French from linguistics, education and history.

A text can be defined as a unit where sentences and larger segments are linked to each other by semantic relations (Halliday, 1985: 318, Halliday & Hasan, 1989: 48). Therefore, a text is also always organised according to some criteria. Text sequences offer one possibility for this task, and they are often defined as structures organising text to equivalent or parallel items (Jackiewicz, 2005, Adam & Revaz, 1989). Many studies (*inter alia* Péry-Woodley, 2000:137-

142, Virbel, 1999, Laippala, 2008) have nevertheless shown that in practice sequence items can be very heterogeneous. In the present study, one of the features indicating this heterogeneity is the frequency of unmarked items, i.e. items that are not introduced by a marker of addition or order, such as *first / premièrement, in addition / de plus*, etc. In the corpus, *partially unmarked* sequences that are composed of both marked and unmarked items are more frequent than sequences where all the items are explicitly signalled. This unmarking can to some extent be explained by the semantic status of text structure: even though semantic relations between sentences can be signalled by these markers, the relations can as well be left unmarked. The reasons for the marking of some relations and the unmarking of others, however, need further studies.

This paper is a part of a larger work on text sequences, their markings and their role in text organisation. The aim of this paper is to present first steps in the analysis of partially unmarked sequences and the unmarked items in them. These sequences form an interesting sequence subtype because of their variation from the standard ones. In addition, they offer a good insight for the study of unmarking. Does the unmarking concentrate on specific positions in the sequences? And, more generally, how do partially unmarked sequences differ from other text sequences? The following example is an illustration of these sequences. The first item is considered as unmarked since it is not explicitly signalled by a marker of addition or order. Explicit markers are in bold.

(1) 1. *Les archives de trois établissements […] ont fait l'objet d'un dépouillement exhaustif.*

2. *Nous avons* ***aussi*** *analysé les Mémoires et souvenirs de 58 anciens élèves. […]*

3. ***S'ajouten****t à ces Mémoires une dizaine de biographies d'hommes publics qui consacrent quelques pages aux études collégiales. […]*

1. The archives of three schools […] have been studied carefully.

2. We have **also** analysed the published and unpublished memoirs and memories […]

3. To these memories **can be added** a dozen biographies by politicians who devote a couple of pages to their college years. […]

The article starts by discussing previous studies concerning text sequences and by defining the structures for this study. The section 3 presents the corpus and the method of the study. In the section 4, partially unmarked sequences and sequences with marked items only are compared. Finally, the section 5 studies partially unmarked sequences and on the positions of unmarked items in them.

2 Text sequences in previous work and in this study

Early work on text sequences concentrate on the series of markers used in them. For instance Turco and Coltier (1988) and Adam and Revaz (1989) study different markers of enumeration, such as *first, second, finally / premièrement, deuxièmement, enfin* and *on one hand – on the other / d'une part – d'autre part.* These articles can be seen as studies on markers of cohesion that are used to signal semantic relations between different parts of text

(Halliday & Hasan, 1976). These markers are also linguistic signs of coherence, a property of texts that can be defined as a principle according to which texts are understood as a whole (Charolles, 1994, 1997) or as a semantic property of them (Halliday, 1985). However, markers of cohesion only signal the relation and guide readers through text (e.g. Charolles, 1994, 1997). A text can also form a well-organised, coherent entity without any surface markers. In this case, the reader needs to interpret the relations between text segments based on situational, textual and propositional knowledge (see, e.g. Halliday and Hasan, 1989, Charolles, 1994, 1997). This is the case also for unmarked items in text sequences.

More recent work on text sequences observe them as text-organising structures that are signalled in the text by different means. Jackiewicz (2005) et Jackiewicz et Minel (2003) form a basic description of text sequences and the marker patterns used to introduce them with a long-term goal of applying the results to information extraction. Luc and Virbel (2001) and Luc et al. (2002) study text sequences in the framework of MAT, a model that includes also typographical item markers, such as dashes and numbers, in the analysis. Typographical markers are also included in the analysis in the Annodis project (e.g. Ho Dac et al., 2009) aiming at an XML-annotated corpus covering text sequences and their items.

In addition to the inclusion or exclusion of typographical markers in the analysis, another aspect dividing the work on text sequences concerns the status of order between the sequence items. For instance Adam & Revaz (1989:66) note that '*an enumeration is not controlled by any order*'[1] whereas Jackiewicz (2005:107) counts order as one of the main characteristics of the structures. In the Annodis project (Ho Dac et al., 2009), the sequence items do not need to be ordered. In this work, text sequence items are considered as at least partially ordered. In addition, typographical markers are left outside the analysis. The main reason for these choices is the restriction of the study to a subtype of enumerations which enables a more detailed analysis. At least partially ordered text sequences may also be more easily identifiable than enumerations without any order. For instance, while enumerations without order can be introduced by an unlimited number of ways, such as spatial adverbs (*In France…In Spain…)* or the repetition of similar lexico-syntactic patterns (*Research articles in history… Linguistics articles…),* explicit means to express order between the items are more limited. This becomes an advantage especially in the analysis of heterogeneous sequences where item marking can vary significantly or be absent.

3 Corpus and methods

The corpus of the study consists of altogether 90 reviewed and published research articles (794 378 words) in linguistics, education and history. The articles have been published between 2000 and 2005 in journals such as *Revue Romane*, *Marges Linguistiques, Recherches et Educations* and *Cahiers d'histoire.*

The sequences, sequence items and their markings are annotated manually in the corpus with the XML annotation tool Callisto (mitre.callisto.org). Even though some text sequences could be retrieved automatically from texts, hand-annotation is the only reliable method to collect all sequences of the corpus. In addition, it allows a comprehensive study of the sequences as

[1] My translation.

all sequence markings can be included in the analysis and not only the ones previously known.

Text sequences need to be defined in detail in order to guarantee reliable and accurate results on their use. This is of particular importance in a quantitative study, as the sequences that are included and the ones that are excluded need to be formally distinguished. Since text sequences can take diverse forms, Porhiel (2007:109) suggests that their definition should be based on elements they are composed of and not on common features different sequences may have. The present study follows this principle and defines text sequences by establishing criteria on the surface marking of sequence items. As the sequence items need to be at least partially ordered, at least some of the item markers also need to indicate addition or order. This forms the basis of the distinction of unmarked and marked items as well as of the required surface features. Nevertheless, since markers used to signal sequence items are very polysemic and since the structuring of a text as a sequence does not always depend on any surface structures, it must be admitted that no surface features can capture all and nothing but all sequences in the corpus. The criteria applied in the annotation are listed in the following.

1. The minimum length of the sequence is two sentences.

2. At least one of the sequence items needs to be signalled explicitly by a marker of addition or order. If the marker signals only addition (such as *the following* or *in addition*) and not exact place of the item in the sequence (such as *the first, finally or thirdly*), the sequence needs to be preceded by a *prediction* or followed by an explicit *closure*. For instance, a sequence where the only explicit marker is *first* or *finally* is included in the corpus. However, if the only explicit marker is *then*, the sequence needs to be preceded by a prediction (such as *this sequence consists of a number of items*) or followed by a closure (such as *to conclude*) to be analysed.

3. If all of the items are unmarked, the sequence is included in the analysis if it is preceded by an **exact** prediction or followed by an **exact** closure that defines the number of the items in the sequence: *In this sequence, there are two parts…*or *These two parts…*

4 Partially unmarked sequences in general: long and vague

In the corpus, there are 311 sequences with marked items only, and 335 partially unmarked sequences. The first distinguishing feature of the partially unmarked sequences is their length: compared to sequences with marked items only, partially unmarked sequences are clearly the longest. The mean and median lengths of the two sequence types are explained in Table 1.

Sequence type	Sequence length mean	Sequence length median
All partially unmarked sequences	391 words	164 words
Partially unmarked sequence **excluding** section sequences	249	154
All sequences with marked items only	265	114
Sequences with marked items only **excluding** section sequences	181	110

Table1. Differences between the lengths of partially unmarked and marked sequences.

The high average length of partially unmarked sequences can to a degree be explained by the frequency of section sequences, i.e. sequences where the items correspond to sections with headings. These sequences are naturally very long and are the most frequent in this sequence type, with 21 out of the 35 section sequences of the whole corpus being partially unmarked. Despite this, partially unmarked sequences are also the longest when section sequences are excluded from the analysis[2]. Therefore, it would seem that longer text sequences are less explicitly marked than shorter ones. Example 2 below illustrates section sequences. Headings are underlined and the explicit markers in bold.

(2) 0. *Dans la section suivante, nous proposons* ***une nouvelle batterie de tests,*** *[…].*

3.3. Seconds diagnostics : VPE vs anaphore nulle

Déterminer si on est en présence d'une ellipse, dans le sens d'un matériel syntaxique/sémantique qui a été supprimée, ou d'une anaphore (nulle) n'est pas chose facile. […].

1. *3.3.1. Recyclage d'un antécédent*

Une différence fondamentale entre ellipse et anaphore (nulle) concerne la capacité de « recycler » un matériel absent dans l'antécédent. […]

2. *3.3.2. Les constructions ACD*

Comme nous l'avons déjà souligné plus haut § 2 (cf. Abeillé, 1991), l'ellipse modale partage avec la VPE anglaise, le fait d'être licite dans les constructions relatives […]

3. *3.3.3. Le Pseudo-gapping*

Le dernier test *est un peu moins direct, puisqu'il se fonde sur […]*

Closure:*Les résultats* ***des tests ci-dessus sont*** *résumés dans le tableau suivant :*

0: In the following section, **I propose a new set of tests** […].

3.3.Second diagnosis: verb-phrase ellipsis vs. null anaphora

It is not easy to distinguish between ellipsis and null anaphora, the first in the sense of an omission of syntactic or semantic material. […]

[2] The difference between the lengths of marked and partially unmarked sequences when section sequences are excluded is also statistically significant with the Kruskal Wallis test with a p-value = 1.878e-07.

1. 3.3.1. Antecedent recycling

   A fundamental distinction between ellipsis and (null) anaphora concerns the capacity to "recycle" material that is absent in the antecedent. […]

2. 3.3.2. Antecedent Contained Deletion

   As I have emphasised above in $ 2 (see Abeillé, 1991), modal ellipsis has in common with the English verb-phrase ellipsis the capacity to be possible in […]

3. 3.3.3 Pseudo-gapping

   The last test is somewhat less direct, because it is based on an […]

   Closure: The **results of these tests above** are presented in the following table: […]

The section sequence in example 2 is composed of three items corresponding to whole sections. In addition, the sequence is preceded by a prediction that actually starts already in the end of the previous section 3.2 and continues in the introduction of the section 3.3. The prediction also announces the topic of the following text and indicates that it is structured as *a set of tests*. Finally, the sequence ends with a closure that refers to the sequence items as a group, *these tests,* and indicates the ending of the sequence.

The only explicit item marker in the actual sequence is *the last test / le dernier test* in the third item. The sequence structure is also declared to some extent by the prediction and the closure. In addition to these explicit markers of enumeration, section sequences have the advantage of being also delimited by the headings and the section boundaries that naturally segment the text (Jacques and Rebeyrolle, 2006) and thus contribute to the explicit marking of the sequence. However, these signs differ from the primary markers of enumeration, i.e. those signalling addition and order as they do not indicate addition or the position of the item in the sequence and as they most often do not function as item markers. Therefore, headings function rather as complementary signs than explicit signals of sequence items.

In addition to the sequence lengths, another distinctive feature of partially unmarked sequences is the frequency of sequences with only vague markers, i.e. additive markers that do not specify the order or the exact position of the item in the sequence (*in addition / de plus)*. Also vague predictions and closures (such as the closure of the example 2) not indicating the number of sequence items but e.g. only announcing its beginning or ending are more frequent in partially unmarked sequences. Table 2 presents the exact frequencies of these properties.

	Item number **not** indicated before or after (exact number)	Vague markers only (exact number)
Partially unmarked sequences	83% (279 sequences)	53% (177 sequences)
Marked sequences	65% (202 sequences)s	2% (6 sequences)

Table 2.

As can be seen in Table 2, exact predictions and closures are more frequent in marked sequences than in partially unmarked sequences[3]. Moreover, more than 50% of the partially

[3] The difference is also statistically significant with a p-value = 1.537e-07 with the X2-test.

unmarked sequences are introduced by vague markers only, whereas marked sequences with only vague markers are almost nonexistent. As partially unmarked sequences also include unmarked items, they seem to organise text to more vague segments which are less explicitly ordered and resemble more additive structures than marked sequences. In other words, less explicit and vaguer marking of text organisation seem to come together. Nevertheless, partially unmarked sequences can as well structure text to well defined and clearly ordered items – this is simply less frequent.

The difference between partially unmarked sequences with no information about the number of items and the ones where this is defined can be considered in examples 1 and 3. In example 1, only vague markers are used, which makes the resulting sequence more of an additive structure. In example 3, in contrast, the number of items is indicated both before and after the sequence, and also the last item of the sequence is marked by *finally / finalement*, signalling necessarily the last item. This makes the sequence structure and the order of the items more explicit and diminishes the effort required by the reader to interpret the text.

(3) 0. *Banks (1989, p. 192) identifie* ***quatre approches*** *du multiculturalisme, relevant de différents niveaux de complexité.*

1. *Au niveau le plus superficiel, les approches « contributionnistes » se limitent […].*

2. *Les approches « additives » ajoutent des contenus, des concepts, des thèmes […].*

3. *Situées à un niveau plus profond, les approches « transformationnelles » […].*

4. ***Finalemen****t, les approches « d'action sociale » impliquent les membres des groupes culturels […].*

Clo.***Ces quatre types d'approches*** *guideront l'analyse de la philosophie de l'éducation multiculturelle […].*

0. Banks (1989, p. 192) identifies **four approaches** to multiculturalism with different degrees of complexity.

1. On the most artificial level, the contributional approaches confine themselves […].

2. The "additional" approaches add meaning, concepts, themes […].

3. At a deeper level, "transformational" approaches aim at changing […].

4. **Finally**, the "social engagement" approaches involve members of multicultural groups […].

Clo. **These four approach types** will guide the analysis of the philosophy behind multicultural education […].

In addition to the explicit markers of enumeration, all the sequence items in example 3 include repetitions of the words *approach / approche*. As discussed briefly in section 2, this creates between the items parallelism, typically related to text sequences. However, as in section headings, repetition does not signal addition or order and can also be misleading as an item marker. Therefore, repetition and other forms of syntactic or lexical parallelism are also considered as complementary markers of enumeration.

5 Partially unmarked sequences in more detail: positions of the unmarked item(s)

The position of the unmarked item(s) in the sequence is an interesting issue that could clarify reasons for the unmarking of sequence items. For instance, a prediction defining the number of sequence items (e.g. *This sequences has two items…)* could explain the unmarking of the first item. The explicit marking of the first item and the unmarking of the others, in contrast, could indicate that the first item is the most important and therefore signalled explicitly, or that it is presenting crucial information that needs to be given to the reader urgently and before the other items (Enkvist, 1989, Virtanen, 2004).

To analyse the position of unmarked items, the sequences were divided to two subgroups according to the marking of the first item[4]. In first the group, the first item is not explicitly signalled, whereas in the second group it is marked explicitly. The rest of the items can be either marked or unmarked, as long as at least one of the items is explicitly marked and one is left unmarked. Table 2 describes these groups and their frequencies as well as presents the frequencies of predictions and vague markers. These figures are discussed later in this section.

Position of the unmarked item(s)	Frequency	Sequences with a prediction	Vague markers only (exact number)
First item **unmarked**	274	28% (76 sequences)	61% (167)
First item **marked**	61	44% (27)	16% (10)
All partially unmarked sequences	335	31% (103)	53% (177)

Table 3. Frequencies of unmarked items in different positions in the sequence.

5.1 Sequences with the first item unmarked

First of all, the results show that the positioning of the unmarked items in the sequence is by no means arbitrary. Clearly the most frequent partially unmarked sequences are the ones with the first item unmarked. A typical feature of these sequences seems to be the frequency of vague markers, such as *in addition / de plus* and *another example / un autre exemple*: they are the only markers used in 61% of these sequences, whereas in the other partially unmarked sequences they are the only markers used in only 16% of the sequences (see Table 3).

As for predictions, as opposed to what was presumed in 5.0, it seems that the unmarking of the first item would not be caused by a prediction already defining the structuring of the sequence, since only 28% of the sequences are preceded by a prediction. In fact, the results suggest that predictions may even be slightly less frequent in this sequence type than in others[5] (see Table 3). It would therefore seem that additive marking is simply typical of sequences with the first item unmarked. This was already said to be regular for partially unmarked sequences in general, but it also seems to be concentrated on sequences with the

[4] As the limits of this article are restricted, a more detailed analysis of the subgroups will be done in another study.

first item unmarked. Example 1 illustrates these structures very well: it only includes additive markers and is not preceded by a prediction.

5.2 Sequences with the first item marked

Compared to sequences with the first item unmarked, sequences with the first item marked are considerably less frequent with only 61 occurrences. In addition, they are typically introduced by exact markers indicating order of the item in the list: only 16% of these sequences include only vague markers (see Table 3).

The sequence type includes two typical structures. In the first structure, the first item is generally the only explicitly signalled item. This could suggest that these sequences emphasise the first item as the most important and leave the others to background.

(4) 1. *le mandat fut exploité* ***d'abord*** *par les lecteurs attentifs des Compilations que furent les canonistes qui édifièrent une théorie de la délégation particulièrement en matière de justice.*

2. *À leur tour les civilistes bénéficièrent de cette réflexion de leurs rivaux.*

1. the mandate was **first** exploited by the alert readers of the Compilations, the canonists, who constructed a theory […].

2. Specialists in the civil law, in turn, benefited from these reflections of their rivals.

In example 4, the first item, signalled by *first / d'abord*, refers to the canonists having been the first to exploit the mandate. The second item, also identifiable by a *in turn / à leur tour*, talks about the civilists, who then could benefit from the reflections of the canonists. Therefore, in this sequence the order between the items appears crucial, as the main reason for the positioning of the second item seems to be the succession of the items in time. If the first item is also considered as the most important, it is not the first criteria for the sequence marking as was presumed. This is also the case in many other sequences with the first item explicitly introduced. However, further studies are still needed to decide whether all the sequences of this type function similarly.

The second typical structure of sequences with the first item marked is the one where also the last item is explicitly signalled. These sequences are always composed of at least three items, the middle one or ones being unmarked. The most frequently, one of the markers is exact and one vague. Example 6 below illustrates these sequences.

(6) 1. ***Au premier plan*** *des destinations de ces exportations, et loin devant l'Amérique Latine, viennent les pays industrialisés ou en voie d'industrialisation. […]*

2. *Les exportations vers le reste de l'Europe et la Russie, vers l'Amérique Latine et la façade atlantique des États-Unis s'élèvent […].*

3. *L'Afrique du Nord et le Sénégal apparaissent au début de la Monarchie […]*

4. ***Enfin****, on trouve en dernière position et dès les années 1830 les Indes et 30 ans plus tard la Chine et le Japon […].*

[5] The X2-test does not entirely confirm the difference between the frequencies of predictions in the sequences with the first item unmarked and marked. The p-value = 0.01749.

1. **In the foreground** of the destinations of these exportations, far before Latin America, are industrial and industrialising countries. […]
2. Exportations to the rest of the Europe and to Russia, to Latin America and to the east cost of the USA grow […].
3. North Africa and Senegal spring up in the beginning of the July Monarchy. […].
4. **Finally**, in the last position since 1830 India and 30 years later China and […].

The sequence in example 6 is an enumeration of destinations for clothing exportation in the 1800 century. A distinguishing feature of this sequence is the parallelism of the items: they all refer to countries or continents. The explicit signalling of the first and last items seems to also emphasise their position in the sequence and help considerably the interpretation of the sequence boundaries. As a whole, compared to example 5, this sequence appears noticeably more explicitly structured.

6 To conclude

The aim of this article was to explore the marking of text organisation by text sequences and, in particular, to concentrate on partially unmarked text sequences and the unmarked items in them. The first part of the analysis observed partially unmarked text sequences in comparison to sequences with marked items only. According to the results, partially unmarked sequences are long and often introduced by only vague markers indicating addition, when compared to sequences where all the items are explicitly marked. It would therefore seem that the marking of text organisation becomes less explicit and vaguer at a more global level when the sequence length increases. The relative rarity of exact predictions and closures in these sequences could also support this.

The second part of the article concentrated on a more detailed study of partially unmarked sequences by dividing them to two groups according to the marking or unmarking of the first item. The results show that the positioning of the unmarked items is not arbitrary. By far the most frequent were the sequences with the first item unmarked. As opposed to what was presumed, the unmarking of the first item generally does not seem to be provoked by a prediction already announcing the beginning of the sequence. Instead, these sequences seem to be used to organise additive structures with less exact markings. Unlike these sequences, the ones with the first item marked are often signalled by exact markers indicating the exact position of the item in the list. Sequences only introduced by vague markers are in this category less frequent. Moreover, in these sequences, the first item is often the only one signalled explicitly. According to the study, this is frequently caused by the stressing of the linear succession of the items in time.

To conclude, the results suggest that partially unmarked sequences are not a uniform group but include several structure types with different means or models for text organisation. These differences can as well explain some of the differences in the use of exact or vague markers or the unmarking or marking of certain sequence items. However, especially the unmarking and the reasons behind it raise still many open questions that this paper could not answer. Why some sequence items are left unmarked? How the unmarked items are identified?

References

ADAM M. & REVAZ F. (1989). Aspects de la structuration du texte descriptif : les marques d'énumération et de reformulation. *Langue française* 81, 59-98.

CHAROLLES M. (1994). Cohésion, cohérence et pertinence du discours. *Travaux de linguistique 29,* 125-151.

CHAROLLES M. (1997). Encadrement du discours: univers, champs, domaines et espaces. *Cahiers de Recherche Linguistique* 6, 1-73.

ENKVIST N.E. (1989). Connexity, interpretability, universes of discourse, and text worlds. In: ALLÉN S. (ed), *Possible words in Humanities, Arts and Sciences: Proceedings of Nobel symposium* 65, 162-186. Berlin / New York:Walter de Gruyter.

HALLIDAY M.A.K. (1985). *An Introduction to Functional Grammar*. USA, Australia:Edward Arnold.

HALLIDAY M.A.K., HASAN R. (1976). *Cohesion in English* London: Longman.

HALLIDAY M.A.K., HASAN R. (1989). *Language, context and text: aspects of language in a social-semiotic perspective.* Oxford University Press.

HO-DAC L-M., FABRE C., PÉRY-WOODLEY M-P., REBEYROLLE J. (2009). *Corpus annotation of macro-discourse structures. 1st International conference on corpus linguistics*, 7-9 may 2009, Murcia, Spain.

JACKIEWICZ A. (2005). Les séries linéaires dans le discours. *Langue française* 148, 95-110.

JACKIEWICZ A., MINEL J. (2003). L'identification des structures discursives engendrées par les cadres organisationnels *Actes de la 10e Conférence Traitement Automatique du Langage Naturel (TALN 2003), Batz-sur-Mer*, 155-164.

JACQUES M.-P., REBEYROLLE J. (2006). Titre et structuration des documents. *Schédae. Prépublications de l'Université de Caen.* Caen : Presses universitaires de Caen, 1-12.

LAIPPALA V. (2008). Nature des marqueurs des séries linéaires. In: DURAND ET AL. (Eds.) *Actes du Congrès mondial de linguistique française 2008*. Paris: Jouve.

LUC C., VIRBEL J. (2001). Le modèle d'architecture textuelle : fondements et expérimentations. *Verbum* 23, 1, 103-123.

LUC C., MOJAHID M., VIRBEL J. (2002). Le modèle d'architecture de texte. *Actes de ISLsp 2002*. Toulouse : Prescot.

PÉRY-WOODLEY M-P. (2000). *Une pragmatique à fleur de texte : approche en corpus de l'organisation textuelle.* Université de Toulouse-LeMirail, ERSS.

PORHIEL S. (2007). Les structures énumératives à deux temps. *Revue Romane* 42(1), 103-135.

TURCO G., COLTIER D. (1988). Des agents doubles de l'organisation textuelle, les marqueurs d'intégration linéaire. *Pratiques* 57, 57-79.

VIRBEL J. (1999). *Structures textuelles - planches, fascicule 1 : Énumérations.* Rapport IRIT, Toulouse.

VIRTANEN T. (2004). Point of departure: cognitive aspects of sentence-initial adverbs. In: VIRTANEN T. (ed.), *Trends in Linguistics. Approaches to cognition through text and discourse*. Berlin: Mouton.

Role of lexico-syntactic and prosodic cues in spoken comprehension of enumerations in sighted and blind adults[1]

Stéphanie Giraud (1), Pierre Thérouanne (2)

(1) LudoTIC, Nice & Laboratoire de Psychologie Cognitive et Sociale,
Université de Nice – Sophia Antipolis, France
stephanie@ludo-tic.com
(2) Laboratoire de Psychologie Cognitive et Sociale, Université de Nice –
Sophia Antipolis, France
therouan@unice.fr

Abstract

Signaling text organization by different ways may improve comprehension. Two experiments on sighted and blind people studied the benefit provided by signals for spoken language comprehension of expository texts including an enumeration. In addition, these studies tested whether the benefit provided by signals was more important for deep comprehension than for the surface structure of texts. Results showed that comprehension was facilitated when texts were presented with prosodic cues. Moreover, lexico-syntactic signals facilitated comprehension when it required understanding specific semantic relationships between co-enumerated items. However, benefit provided by these signals was restricted to blind participants. Results are discussed in terms of expertise and suggest that signaling should improve access to information for blind people.

Keywords: Comprehension, Spoken language, Text format, Enumeration, Blindness

1 Introduction

Information and communication technologies are widely used in everyday life. In one hand, they create a new source of exclusion for people with visual impairment. On the other hand, these technologies potentially fill in their disability by offering access to information with computer interfaces like text-to-speech synthesizers (TTS). Information about the structure of

[1] We thank Léa Pasqualloti and Clément Filippi for their help in designing study and data collection, and the two anonymous reviewers for their helpful comments.

texts conveyed by visual signals of written text should be preserved when texts are oralized by TTS.

1.1 Role of signals in text comprehension

Signaling text organization can be achieved in many ways in written language, with discursive, lexico-syntactic signals (e.g., conjunction), typographic signals (e.g., dash, numbering) and dispositional signals (e.g., horizontal and vertical spacing). According the Textual Architecture Model (Virbel, 1985), signals are realizations of metasentences (e.g., “the first part of the text is…”) describing the elements of the written text itself. In spoken language, intonation, melody, pauses and emphasis may also convey information about text structure.

There is a growing body of evidence that signaling text organization leads to a better comprehension. Lorch and Lorch (1996) showed that headings improve global comprehension of texts. Likewise, Lorch et al. (2001) showed that signals effects on the recall of information result from modifications of text representation are responsible effects of signals. Besides, signals effects varied according expertise level; the benefit provided by signals was more important for novice readers than for expert readers. Schmid and Baccino (2002) showed that formatting text with dispositional signals helped readers to identify perspective shift in narrative texts. Lemarié, Eyrolle and Cellier (2006) showed that discursive and prosodic cues improve the comprehension of restaurant menu: these cues helped them to develop adequate representation of the oralized texts. Finally, Lemarié, Lorch, Eyrolle and Virbel (2008) proposed a model that integrates linguistic and cognitive analyses of signals: SARA (Signal Available Relevant Accessible information). They characterized signals and how signals achieve their effects depending on the availability, relevance and accessibility of information. These authors showed that the magnitude of the effect of a signaling device will increase as its task relevance increases. Signals effects also depend on capacities, knowledge and goals of reader. In addition, discursive signals should be used when the author wants to increase the likelihood that the signaled content will be carefully processed.

However, cognitive processes underlying signals effects need to be investigated. Text comprehension is generally conceived as the successive construction of three levels of representation (Van Dijk, Kintsch, 1983): the *surface structure* consisted of the original words in the text, the *text base* which is the semantic content of the text, and the *situation model* corresponding to the situation described by the text. Different claims have been made about the representational level that would benefit from signaling text organization (see Lemarié et al., 2006). Indeed, Schmid and Baccino (2002) and Maurel et al. (2003) stated that signals leave no trace in the surface structure. Maurel et al. (2003) suggested that signals enable to develop a high level representation through a deep processing, whereas Schmid and Baccino (2002) proposed a fourth level of representation: the *organizational or spatial level*.

1.2 Enumeration

Enumerative structure consists of an introducer, an enumeration listing at least two items, and an optional conclusion (Luc, 2001). The introducer announces the enumeration and can be complete when indicating the number of co-listed items. Furthermore, Luc (2001) distinguished two kinds of enumerations: paradigmatic enumerations listing items that are functionally equivalent, and syntagmatic enumerations exhibiting syntactic or semantic dependence between items.

1.3 The present research

Previous study we conducted showed that reading time of paradigmatic enumerations presented in isolation was shorter when typographic cues ("-") signaled the co-enumerated items than when lexico-syntactic cues (e.g., "first of all", "then", "finally") were used as signals. Nonetheless, no difference was obtained on comprehension questions. A second experiment did not show any benefit provided by lexico-syntactic cues in spoken language. We made the assumption that texts more complex to process would exhibit a greater benefit from cues. Thus, paradigmatic enumerations were included in larger texts in Experiment 1, and Experiment 2 studied syntagmatic enumerations with semantic dependence between items. Prosodic cues (pauses) and level of comprehension question were also manipulated in Experiment 1.

People with sighted and blind people participated to the two experiments. Previous studies showed that individuals with visual impairment point out better processing of sounds and spoken language than individuals with sight (e.g., Edmonds, Pring, 2006; Röder, Rösler, 2003). Then, we hypothesized that the benefit from cues would be greater for sighted people than for blind people.

2 Experiment 1

2.1 Participants

All participants in the two experiments were French native speakers with no reported hearing difficulties. In Experiment 1, sighted participants were 24 undergraduate and graduate students following a Psychology course, 22 females and 2 males from 18 to 26 years old (median = 22), with normal or corrected sight. Blind participants were 12 members of Valentin Haüy Association and Civil blinds Association. They were 8 females and 4 males aged from 33 to 69 years old (median = 50.5). They were congenitally blind or were blind for at least two years. Their education degree ranged from secondary school to graduate studies.

2.2 Materials

Experimental stimuli consisted of 48 expository texts including a paradigmatic enumeration with 3 or 4 items. Texts were adapted from textbooks for secondary schools and electronic encyclopedias (e.g., Wikipedia) and covered numerous knowledge domains. Enumeration was preceded by an introduction and/or followed by a conclusion. Each enumeration has been created in two versions: an interpretative version and a restricted version (see Figure 1). In the interpretative version, enumeration reconstructs the architectural intentions of author by interpretation; the introducer indicated the number of co-listed items and each item was preceded by a lexico-syntactic cue naming the item category. In the restricted version, no textual cue was given; the introducer was incomplete and lexico-syntactic cues for items were absent. Auditory stimuli were generated with a speech synthesizer (Infovox Desktop, Acapela) using a female voice chosen for its clarity. Prosody was also manipulated. In the version with prosodic cues, the pause ending a sentence lasted 600 ms., the pause ending the introducer lasted 400 ms., and pauses between co-enumerated items lasted 200 ms. In the version with no prosodic cues, all pauses lasted 600 ms.

Text	
Interpretative version	Restricted version
Le pingouin est un oiseau noir et blanc de la famille des alcidés, appelé également petit pingouin ou pingouin torda. Les trois caractéristiques différenciant le pingouin du manchot sont énumérées ci-après : la première caractéristique est la zone où il vit, située entre l'Océan Arctique et la Bretagne ; la deuxième caractéristique est sa capacité à voler au dessus de l'eau ; la troisième et dernière caractéristique est sa capacité à plonger en apnée limitée à deux minutes au maximum.	Le pingouin est un oiseau noir et blanc de la famille des alcidés, appelé également petit pingouin ou pingouin torda. Les caractéristiques différenciant le pingouin du manchot sont : la zone où il vit, située entre l'Océan Arctique et la Bretagne ; sa capacité à voler au dessus de l'eau ; sa capacité à plonger en apnée limitée à deux minutes au maximum.

Question	
Surface structure	Situation model
Quel mot était dans le texte original ? sa _________ à voler au dessus de l'eau. A. facilité B. compétence C. capacité D. faculté	Où vivent les pingouins ? A. Dans les deux hémisphères B. Uniquement au pôle Nord C. Dans l'hémisphère Nord D. Dans l'hémisphère Sud

Figure 1: Example of Texts and Questions for Experiment 1.

For each text, two questions were elaborated according Daniel and Raney (2007) principles. One correct answer and three alternatives were presented for each question. For the surface structure question, participants had to fill the missing word. For the situation model question, participants had to produce an inference to choose the correct answer. Questions were divided equitably between the introducer and the co-listed items to avoid focusing attention on specific parts of texts. Text version, Prosody, and Question were within-subjects factors.

Six texts without any enumeration were used as fillers, with questions addressing the text base level. Four texts were used for the practice phase, one in interpretative condition, one in restricted condition and two without any enumeration.

2.3 Procedure

Experiment was conducted in laboratory for sighted participants and in a non isolated room outside the laboratory for blind participants. Once the practice phase completed, the 48 experimental and 6 filler trials were presented in a random order. Participants were seated in front of a computer screen and listened texts presented with headphones. Each text was followed by a question. The question and the 4 answers were presented on the computer screen for sighted participants and they were presented with headphones for blind participants[2]. They were instructed to indicate the correct answer by pressing one of four keyboard keys. Then, they pressed the spacebar to start the next trial.

2.4 Results

Correct response percentage was higher for sighted participants (M = 63%) than for blind participants (M = 45%), $F_1(1, 34) = 21.37$, $p < .01$; $F_2(1, 47) = 24.16$, $p < .01$. Correct response percentage was higher with prosodic cues (M = 58%) than with no prosodic cue (M = 52%), although this difference did not reach significance in the subjects' analysis, $F_1(1, 34) = 3.93$, $p = .06$; $F_2(1, 47) = 4.11$, $p < .05$. The four experimental factors - Group, Text version, Prosody, and Question - interacted in the subjects' analysis, $F_1(1, 34) = 6.17$, $p < .05$; $F_2(1, 47) = 3.72$, $p = .06$. No other significant effect was obtained.

		Text version				
		Interpretative		Restricted		
Group	Question	prosodic cues	no cues	prosodic cues	no cues	Mean
Sighted	Surface structure	11007 (65%)	10566 (64%)	11700 (72%)	10741 (58%)	11004 (65%)
	Situation model	17427 (67%)	15633 (56%)	16175 (60%)	16967 (65%)	15114 (62%)
	Mean	14217 (66%)	13100 (60%)	13938 (66%)	13854 (62%)	
Blind	Surface structure	13618 (46%)	15999 (39%)	17515 (50%)	17379 (50%)	16127 (46%)
	Situation model	15563 (44%)	12612 (43%)	20454 (47%)	16365 (39%)	16248 (43%)
	Mean	14590 (45%)	14305 (41%)	18984 (49%)	16872 (45%)	

Table 1: Mean correct response latencies (ms) and percentage of correct responses (in parentheses) as a function of Group, Text version, Prosody and Question. Experiment 1.

[2] On one hand, procedure should be the same for the two groups. On the other hand, questions were used to test text comprehension itself, and not initial knowledge of participants. Thus, questions were presented in modality which was more convenient for each group, resulting in a variation of procedure between sighted and blind groups. For this reason, any main effect of Group should be interpreted with caution.

Mean correct response latency was shorter in the interpretative version (*M* = 14045) than in the restricted version (*M* = 15824), $F_1(1, 33) = 5.67$, $p < .05$. Furthermore, the interaction between the Group and Text version was significant, $F_1(1, 33) = 4.26$, $p < .05$, showing a greater benefit from interpretative version for blind people than for sighted people.

3 Experiment 2

3.1 Participants

Sighted participants were 33 undergraduate and graduate students following a Psychology course, 23 females and 10 males from 18 to 35 years old (median = 21.5), with normal or corrected sight. Blind participants were 6 members of Valentin Haüy Association and civil blinds Association. They were 2 females and 4 males from 33 to 66 years old (median = 53). They were congenitally blind or were blind for at least two years. Their education degree ranged from secondary school to graduate studies.

3.2 Materials

Experimental stimuli consisted of 42 syntagmatic enumeration adapted from encyclopedias (e.g., Wikipedia) showing semantic dependence between the 3 or 4 co-enumerated items. Each text has been created in three versions (see Figure 2). The interpretative and restricted version followed the principles described in Experiment 1. In the descriptive version, enumeration described explicitly the textual markers, the introducer also indicated the number of co-listed items, and each item was preceded by a lexico-syntactic cue that did not mention the item category. Text version was a within-subjects factor.

Text		
Interpretative version	Restricted version	Descriptive version
Les quatre manipulations nécessaires à l'observation de l'ADN de l'oignon sont énoncées ci-après : la première manipulation est de couper et broyer les morceaux d'oignon dans un mortier contenant une solution d'extraction ; la deuxième manipulation est de filtrer le broyat obtenu et récupérer le filtrat dans un tube à essai ; la troisième manipulation est d'incliner le tube à essais et verser le long de la paroi le même volume d'alcool à brûler ; la quatrième et dernière manipulation est d'ajouter un colorant afin d'observer l'apparition de filaments.	Les manipulations nécessaires à l'observation de l'ADN de l'oignon sont : de couper et broyer les morceaux d'oignon dans un mortier contenant une solution d'extraction ; de filtrer le broyat obtenu et récupérer le filtrat dans un tube à essai ; d'incliner le tube à essais et verser le long de la paroi le même volume d'alcool à brûler ; d'ajouter un colorant afin d'observer l'apparition de filaments.	Voici une liste de quatre éléments concernant les manipulations nécessaires à l'observation de l'ADN de l'oignon : le premier élément est de couper et broyer les morceaux d'oignon dans un mortier contenant une solution d'extraction ; le deuxième élément est de filtrer le broyat obtenu et récupérer le filtrat dans un tube à essai ; le troisième élément est d'incliner le tube à essais et verser le long de la paroi le même volume d'alcool à brûler ; le quatrième élément est d'ajouter un colorant afin d'observer l'apparition de filaments.

Question (Text base level)
Lors d'une manipulation visant à observer de l'ADN d'oignon, quand doit-on utiliser un filtre ?
A. Après avoir versé l'alcool à brûler.
B. Juste après avoir broyé les morceaux d'oignon.
C. Juste avant de broyer les morceaux d'oignon.
D. Juste avant l'ajout d'un colorant.

Figure 2: Exemple of Texts and Questions for Experiment 2.

Twelve additional texts were used as fillers, 3 paradigmatic enumerations, 3 syntagmatic enumerations and 6 texts without any enumeration. The practice phase consisted of 2 syntagmatic enumerations and 2 texts without any enumeration.

For each experimental text, one question about the text base level was elaborated according Daniel and Raney (2007) principles. The question dealt with the semantic (i.e., spatial, temporal or causal) relationship between two items. For the fillers, questions never focused on the semantic relationship between items.

3.3 Procedure

The procedure was identical to the procedure of Experiment 1.

3.4 Results

	Text version			
Group	Descriptive	Interpretative	Restricted	Mean
Sighted	12282 (70%)	12388 (70 %)	12343 (66%)	12338 (69%)
Blind	14601 (45 %)	14873 (37%)	15085 (30%)	14853 (37%)

Table 2: Mean correct response latencies (ms) and Percentage of correct responses (in parentheses) as a function of Group and Text version. Experiment 2.

Correct response percentage was higher for sighted participants than for blind participants, $F_1(1,37) = 38.73$, $p<.001$; $F_2(1,41) = 61.03$, $p<.001$. Main effect of Text version was significant in subject's analysis, $F_1(2,74) = 3.52$, $p<.05$; $F_2 < 1$. Planned comparisons showed more correct responses in the descriptive and interpretative conditions than in the restricted condition, although this difference was not significant in the items' analysis, $F_1(1,37) = 5.22$, $p<.05$; $F_2(1,41) = 1.54$, $p>.10$, whereas the difference between the descriptive condition and the interpretative condition was not significant, $F_1(1,37) = 1.50$, $p>.10$; $F_2 < 1$. Group and Text version did not interacted, $F_1(2,74) = 1.32$, $p > .10$; $F_2 < 1$.

Mean correct response latency was shorter for sighted participants than for blind participants, although this difference did not reach significance, $F_1(1, 37) = 38.34$, $p = .08$. No other effect was significant.

4 Discussion

Experiment 1 did not show any benefit provided by lexico-syntactic cues on comprehension of paradigmatic enumeration. However, Experiment 2 showed such a benefit on syntagmatic enumerations, correct responses being more frequent in interpretative and descriptive versions of texts than in restricted version. This pattern of result confirms our first hypothesis and strongly suggests that lexico-syntactic cues facilitate comprehension when establishing the specific semantic relationships between co-enumerated items is necessary. Experiment 1 also showed that responses were facilitated when texts were presented with prosodic cues, despite the fact that these cues were somewhat subtle variation of pauses duration. However, Experiment 1 failed to show that facilitation provided by signals is greater for situation model than for surface structure.

As blind people rely mainly on spoken language, we predicted that they would less beneficiate from cues than participants with sight. On the contrary, the sighted group was more accurate than the blind group in the two experiments. Moreover, the benefit provided by cues was higher on response latencies for blind participants than for participants with sight in Experiment 1. Although they could be considered as more expert in spoken language activities, blind participants were also older with educational degree largely lesser than participants with sight. Indeed, we failed to match blind and sighted people on main characteristics relevant to this study. We hypothesize that blind participants had fewer background knowledge dealing with the content of the presented texts, making signaling more useful to them (see Lemarié et al, 2008).

Two research directions will be followed. First, given the relations between age, access to knowledge and blindness, it seems that signaling is helpful to improve accessibility of information presented by the mean of spoken language, especially with TTS synthesizers. Second, for a more theoretical purpose, any comparison between blind and sighted participants will require matching on background knowledge and sample size despite the fact that it is difficult to achieve.

References

DANIEL F., RANEY, G. E. (2007). Capturing the effect of a title on multiple levels of comprehension. *Behavior Research Methods* 39, 892-900.

EDMONDS C. J., PRING L. (2006). Generating inferences from written and spoken language: A comparison of children with visual impairment and children with sight. *British Journal of Developmental Psychology* 24, 337-351.

LEMARIÉ J, EYROLLE H., CELLIER J.-M. (2006). Visual signals in text comprehension: How to restore them when oralizing a text via a speech synthesis? *Computers in Human Behavior* 22, 1096-1115.

LEMARIÉ J., LORCH R. F., EYROLLE H., VIRBEL J. (2008). SARA: A text-based and reader-based theory of text signaling. *Educational Psychologist* 43, 1-23.

LORCH R. F., LORCH E. P. (1996). Effects of headings on text recall and summarization. *Contemporary Educational Psychology* 21, 261-278.

LORCH R. F., LORCH E. P., RITCHEY K., MCGOVERN L., COLEMAN D. (2001). Effects of headings on text summarization. *Contemporary Educational Psychology* 26, 171-191.

LUC C. (2001). Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte. Proceedings of *Conférence Traitement Automatique du Langage Naturel (TALN'2001).*

MAUREL F., LEMARIÉ J., VIGOUROUX N. (2003). Evaluation cognitive d'une représentation du texte pour sa présentation multimodale. Proceedings of *Conférence Internationale sur le Document Electronique (CIDE 6).*

RÖDER B., RÖSLER F. (2003). Memory for environmental sounds in sighted, congenitally blind and late blind adults: Evidence for cross-modal compensation. *International Journal of Psychophysiology* 50, 27-39.

SCHMID S., BACCINO T. (2002). Perspective shift and text format: An eye-tracking study. *Current Psychology Letters: Behaviour, Brain & Cognition* 9, 73-87.

VAN DIJK T. A., KINTSCH W. (1983). *Strategies of discourse comprehension*. New York: Academic.

VIRBEL J. (1985). Langage et métalangage dans le texte du point de vue de l'édition en informatique textuelle. *Cahiers de Grammaire* 10, 1-72.

Comparing *because* to *want;* How connectives affect the processing of causal relations.

Anneloes Canestrelli (1), Ted Sanders (2), Pim Mak (3)

(1)Uil- OTS, Utrecht University
A.R.Canestrelli@uu.nl
(2)Utrecht University
T.J.M.Sanders@uu.nl
(3)Utrecht University
W.M.Mak@uu.nl

Abstract. In this paper we investigate the online processing of causal coherence relations between clauses. Starting out from the processing difference between subjective and objective relations, as observed in English (Traxler et al., 1997), we performed two eye-tracking experiments in which we tested these different types of causal relations in Dutch. The results revealed that, as in English, Dutch subjective causal relations take more time to process than their objective counterparts. However, we only observed this asymmetry when subjective and objective causal relations were marked by their prototypical connective, *want* (because) and *omdat* (because) respectively. This finding suggests that the usage pattern of causal connectives affects online processing. We conclude that the subjective-objective distinction has cognitive relevance. We discuss the exact relationship between the nature of causal relations and the properties of connectives.

Keywords: Causality, Subjectivity, Connectives, Processing, Eye-tracking

1. Introduction

Causality is a basic notion in language and cognition. In this paper, we investigate the online processing of causal coherence relations between clauses, focusing on the nature of these relations and how they match the connectives that mark them. Connectives and cue phrases are often regarded as explicit processing instructions of how the parts of a discourse are related to each other (Britton, 1994; Noordman & Vonk, 1997; Sanders & Spooren, 2001). Experimental processing data have shown that readers indeed benefit from the presence of these markers, resulting in faster processing times (Millis & Just,1994; Noordman & Vonk,1997; Sanders & Noordman, 2000) and better comprehension of information (Degand & Sanders, 2002).

Causal coherence relations can be subcategorized in subjective and objective relations (e.g., Pander Maat & Sanders, 2001; Pander Maat & Degand, 2001). The difference between these types of causality can be explained in terms of speaker involvement. In the subjective causal relation 1b, the first clause is a claim on behalf of the speaker on the basis of the real world event in the second clause. As such, the speaker is responsible for constructing the causal relation. This type of relations can be paraphrased as "*From this fact I conclude that"* (Pander Maat & Sanders, 2001). Objective causal relations, such as 1a, do not involve reference to the speaker but relate two states of affairs in the real world. The speaker is not involved in the construction of the causal relation and this sentence can be paraphrased with *"The situation in S2 is the cause of the situation in S1"*.

(1) a. The vase broke because it fell off the table
b. Peter must be sick because he looks pale

Studying on-line text processing, Traxler, Bybee and Pickering (1997) found that subjective causal relations (2b) produced longer reading times on the verb region, *bit him,* compared to objective causal relations (2a).

(2) a. Rick almost died from shock because a poisonous snake *bit him* on the leg
b. Rick was walking in a remote area because a poisonous snake *bit him* on the leg

Traxler et al. propose that the difficulty of these subjective causal relations is explained by the inferences that readers have to make in order to establish the relation. Readers have to infer that the first clause is a speaker's belief rather than a fact in the world. According to the authors subjective causal relations are only difficult when the text has not explicitly made clear that S1 is a claim or hypothesis, and that S2 is the argument in favour of that claim.

Indeed, Traxler, Sanford, Aked & Moxey (1997) demonstrated that the introduction of epistemic markers, such as *I think* and *perhaps* (3), which make clear that the first clause is to be taken as a belief or possibility, cancels the processing difference between objective and subjective causals.

(3) *John thinks* Rick was walking in a remote area because a poisonous snake bit him on the leg

The systematic use of one causal connective over another in specific contexts has led to the hypothesis that there may be a mapping between connectives and domains of use (Stukker & Sanders, 2009). This lexical selection can be interpreted as reflecting the psychological reality of different types of causal relations. In English, both subjective and objective causal relations can be marked by *because* (e.g. Sweetser, 1990; Couper- Kühlen, 1996; Knott & Dale, 1994; Knott & Sanders, 1998). This connective fully matches both types of relations. In Dutch, however, there are two causal connectives that mark backward causality: *Want* and *Omdat*. *Want* is a prototypical marker for subjective claim-argument relations but can be used to mark objective causal relations. *Omdat* is a more prototypical marker for objective causal relations (e.g. Sanders & Spooren, in press.).

Given the cross-linguistic differences between causal connectives, it is relevant to question whether the observed processing asymmetry between subjective and objective causal relations in English can be replicated in Dutch. In addition, we wanted to find out whether the epistemic facilitation effect, due to explicit epistemic markers, is generalizable to more implicit means of subjective marking such as deontic markers and expressions of quantificational aspect. Deontic markers, such as *good*, *bad* and *happy*, convey moral or aesthetic judgements, expressing the speaker's attitude. Expressions of quantificational aspect, such as *always, never, every year*, make clear that the clause is a generalization, and as such may facilitate a subjective interpretation. Two eye-tracking studies were performed to answer these questions.

2. Eye Tracking experiment 1

2.1. Materials and design

The materials consisted of 24 sets of sentences as in (4) of which the a) and b) conditions are Dutch translations of the original objective and subjective items as used by Traxler et al. The c) condition consists of implicitly marked subjective items. These items were designed on the basis of the following characteristics: the first clause was always formulated as a hypothesis, evaluating a characteristic of a person or a person's behaviour, followed by an argument. In contrast to Traxler et al. we did not use epistemic markers, but manipulated the use of deontic markers and expressions of quantificational aspect. All relations are marked by the connective *want*, which can mark both subjective and objective causal relations in Dutch.

(4) a. Rick almost died from shock *want* a poisonous snake bit him on the leg
b. Rick was walking in a remote area *want* a poisonous snake bit him on the leg
c. Rick has been acting carelessly again *want* a poisonous snake bit him on the leg

We made use of a between subjects design to avoid possible interactions between the two subjective conditions. The 24 sets were divided over four lists, according to a Latin square design, so that each list contained only one version of a set. A list either contained original subjective relations or implicitly marked subjective relations combined with objective causals. In addition, 40 filler items were inserted in each list. The experiment started with a practice trial of 5 items.

Line breaks were placed right after the first clause and before the connective. To avoid noise related to the so called return sweep, jumping to the next line (See Cozijn 2000 for more details), we made sure that critical regions were never situated at the end of the first line or at the beginning of the second line. To make sure that participants paid attention to the sentences, verification statements were randomly divided over the stimuli and appeared after 25% of the items. Subjects were informed to respond to these statements by pressing the 'yes' or 'no' button on the button box. Half of the statements were correct, the other half was incorrect. These statements never concerned the causal relations under investigation to avoid that attention was drawn to the purpose of the experiment.

2.2. Participants

Forty-five undergraduate students from Utrecht University participated in the experiment (41 female, mean age 22, 3 age range 18 – 43). All participants were native speakers of Dutch and they received money for their participation.

2.3. Apparatus and Procedure

An SMI Eyelink I head-mounted eye-tracker was used, controlled by FEP software (Veenker, 2006). All subjects were individually tested in a testing boot at the University. Participants were asked to read normally and not to move their heads or blink excessively during the experiment. Each experiment started with a calibration procedure during which subjects had to follow a small circle on the screen moving through 9 positions, followed by a similar validation procedure. This procedure was repeated before each block.

2.4. Results

As found in English, we expect longer fixation times in the original subjective condition (b) compared to the objective condition (a). If the implicit markers of subjectivity manipulated in the present study have a similar function as the epistemic markers reported in Traxler et al., it is expected that there will be no processing asymmetry between objective causals and the implicitly marked subjective condition.

Example 5 illustrates how the items were divided into the regions that were used for our analyses. The regions of interest, where we expected to find effects, were those following the connective: 3, 4 and 5.

(5) [Rick almost died from shock 1]
[because 2] [a poisonous snake 3] [bit him 4] [on the leg 5]

We analyzed the same measures as reported in Traxler et al., namely: First Pass reading times, Right-Bounded Time and Regression Path Duration. First Pass reading time is the durations of all fixations on a word or region before moving into any direction. Right-Bounded Time is the sum of fixations within a region before moving on in a forward direction. Regression Path Duration consists of all fixations on a region before passing on in a forward direction, including rereading of previous material.

A Linear Mixed Effect Regression analysis (LMER) (Baayen, 2008) was performed on the log-transformed data[1]. The idea behind this technique is that a model for the data based on only item and subject variation is compared to a model that takes the manipulation(s) into account. The goodness of fit is assessed with a log-likelihood ratio and informs about which model provides a better explanation for the total variance. The base model was compared with three models to asses the effect of condition (subjective versus objective causal relation), group (original versus implicitly marked) and the interaction between these two. The results

[1] Log transformations were performed on the data to better meet the requirement of a normal distribution, which is a prerequisite for multilevel analyses.

revealed no differences between subjective and objective causal relations in any of the regions.

2.5. Discussion

Contrary to our predictions, Experiment 1 did not reveal differences between objective causals and original subjective causals nor between objective causals and implicitly marked subjective causals. This was unexpected because Traxler and colleagues did find a processing asymmetry between their subjective and objective conditions. Since the present study is a direct translation of their work it is remarkable that we have found no effect at all in our experiment.

Does this mean that subjective causal relations are only difficult to process in English but not in Dutch? Or can the outcome be ascribed to the differences between English and Dutch connectives? Recall that Traxler et al. used the connective *because* to mark all relations in the experiment. As explained in the introduction, this connective is ambiguous with regard to the type of coherence relation it can mark. In our Dutch replication we used the connective *want,* which is also used to mark both types of causal relations. Even though text-linguistic studies cannot provide clear cut distinctions between the domains of use of this connective, it has been shown repeatedly, both in theoretical and corpus work (Degand, 1996; Sanders & Spooren, in press; Verhagen, 2005; Sanders, Sanders & Sweetser, submitted), that *want* is a prototypical marker for subjective claim-argument relations. It is thus possible that the processing instructions provided by *want* are more specific in that they guide the reader towards such a subjective interpretation. This may actually explain why we have found no differences between our conditions: If the processing instruction provided by *want* in exp 1 are in fact something like “here comes an argument to support the previous claim”, similar interpretation processes applied to all conditions and as a result, they may have been equally difficult to process.

The obvious question that arises is: What happens if we match the relations with their prototypical connective, on the basis of corpus results? We therefore conducted a second experiment in which the Dutch connectives were used according to their prototypical usage: *Omdat* marking objective relations and *want* marking subjective relations.

3. Eye Tracking Experiment 2

The materials, design and procedure were exactly the same as in Experiment 1. The only difference between the two studies is that here the objective condition is marked by *omdat* (6a). As explained in the introduction, this connective is a prototypical marker for objective consequence-cause relations.

(6) a. Rick almost died from shock *omdat* a poisonous snake bit him on the leg
b. Rick was walking in a remote area *want* a poisonous snake bit him on the leg
c. Rick has been acting carelessly again *want* a poisonous snake bit him on the leg

3.1. Participants

Forty-seven undergraduate students from Utrecht University, none of which participated in Experiment 1, participated in the experiment (41 female, mean age 23, 6 age range 18 – 45). All participants were native speakers of Dutch and they received money for their participation.

3.2. Results

Regions were defined as in Experiment 1, see example 5. Again, we analyzed the log-transformations of the following measures: First Pass reading times, Right-Bounded Time and Regression Path duration.

First Pass reading times

First pass reading times produced a main effect of condition in region 3 ($X^2(1)=18,49$, $p<0,001$). The segment following the connective induced longer First Pass reading times in all subjective causals compared to objective causals. In addition, we observed a significant interaction between condition and group in this region ($X^2(3)=20,68$, $p<0,001$), indicating that the differences between conditions are different for the two groups. The asymmetry between subjective and objective causals is larger in the group that was presented with the original subjective relations. Post hoc analyses, however, revealed a significant effect for condition in both the 'original' group ($X^2(1)=13,04$, $p<0,001$), and the 'implicitly marked' group ($X^2(1)=5,46$, $p<0,05$).

Right-Bounded Time

Similar effects were found in Right-Bounded Time. We observed a main effect of condition ($X^2(1)=18,11$, $p<0,001$) and an interaction between condition and group ($X^2(3)=19,32$, $p<0,001$). Again, condition was a significant factor in the 'original' group ($X^2(1)=12,66$, $p<0,001$) and in the 'implicitly marked' group ($X^2(1)=6,04$, $p<0,05$).

Regression Path Duration

Again a main effect of condition and a significant interaction between condition and group was observed on the words following the connective in Regression Path Duration (Resp. $X^2(1)=10,80$, $p<0,01$; $X^2(3)=11,30$, $p<0,05$). Condition was a significant factor in the 'original' group ($X^2(1)=6,47$, $p<0,05$) and in the 'implicitly marked' group ($X^2(1)=4,50$, $p<0,05$).

3.3. Discussion

The results of the two experiments indicate that, like in English, Dutch objective causal relations are easier to process than subjective causal relations. This effect was observed in shorter first pass reading times, right-bounded time and regression path durations on the words following the connective. Interestingly, this result can only be replicated when objective causals are marked by *omdat*.

In addition, we have found that implicit marking of subjectivity slightly reduces the processing difficulty of subjective causal relations, resulting in a smaller asymmetry between subjective and objective causals. This demonstrates that our manipulation reduces the processing times in subjective causals, but could not completely solve the processing problem. The effects of the markers of subjectivity in the present study are thus similar to, although weaker than, the epistemic markers reported in Traxler et al. Again, this effect only surfaced when objective causals were marked by *omdat*. In Experiment 1, where all relations are marked by *want,* we could not observe any processing difference at all. So how can we account for the differences between the two experiments?

Given the fact that Experiment 2 replicates the English findings, it is reasonable to believe that there is something different going on in our first experiment. We argue that this difference is due to the properties of the connectives involved.

In the original study, Traxler et al. tested causal relations marked by *because*. This connective is fully compatible with both subjective and objective causal relations. As such, the processing instructions it conveys may be not further specified than marking ‘some sort of causality’. As a result the reader has to rely more on the content of the clauses to establish the relation that holds between them.

The properties of the Dutch connectives used in our study are different: *Want* is considered to be a prototypical marker for claim-argument relations, whereas *omdat* is most commonly used in objective consequence-cause relations. Corpus studies over several text types – varying from newspaper texts to spontaneous conversations - have corroborated this pattern. Given these prototypical usage patterns, it is reasonable to believe that the processing instructions provided by these connectives are derived from the way in which they are commonly used. If *want* guides the interpretation of a relation towards a subjective one, this would mean that our objective relations are in fact processed as subjective. It is then no wonder that we could not observe any differences between the conditions. Indeed, using the more fitting connective *omdat* in the objective condition (Experiment 2) leads to very different results.

These cross-linguistic differences between Dutch and English causal connectives are reflected in the time course of the effects that were found. Recall that Traxler et al. observed the asymmetry between subjective and objective causals in the prefinal region, *bit him* (see example 7). The effects in our Dutch experiment arise earlier, namely immediately after the connective, when the proposition is far from complete. This suggests that indeed we are dealing with two very different processes.

(7) Rick was walking in a remote area because a poisonous snake *bit him* on the leg

But are there alternative explanations for our data? Note that although variation in the first clause is not ideal, because across conditions very different information needs to be integrated with the second clause, it does not account for the results. Recall that we used the same relations in Experiment 1, where we did not find any differences. If the effects are the result of integration differences across conditions, this should also have surfaced in Experiment 1[2].

[2] An explanation on the basis of frequency differences between the connectives does not hold either. According to CELEX the log-frequencies of *omdat* and *want* are 4,617 and 4,539 respectively.

Another alternative explanation may be found in the structural differences between of *want* and *omdat*. Since *want* is a coordinator whereas *omdat* introduces a subordinate clause, it could be argued that the asymmetry in Experiment 2 is not due to the relational difference but merely a result of structural differences. Still, structure on itself cannot account for our data. We demonstrated that the processing asymmetry between *omdat* and *want* clauses is affected by the content of these clauses which has nothing to do with their structure. Also, if structure is the only determining factor, we would have to conclude that Dutch readers have no problems in processing subjective causal relation while English readers experience severe difficulties when reading exactly the same sentences[3].

4. Conclusion

The two experiments reported in the present study demonstrate that objective causals are easier to process than subjective causals, but that this effect depends on the match of the specific connective with the relation. Subjective causals are o*nly* processed slower compared to objective relations when these relations are marked by their prototypical connective, *want* and *omdat* respectively. Furthermore, the implicit markers of subjectivity used in the present study do facilitate the processing of subjective causal relations, although the processing difficulty is not cancelled out entirely.

Overall, the results show that there is an intricate relationship between the nature of the relation on the one hand, and the characteristics of the connective on the other hand. These linguistic markers determine whether or not processing differences between objective and subjective causal relations arise and the position where such effects appear. In case of more ambiguous connectives, such as *because* in English, processing differences surface later than when the connective gives more specific processing instruction, such as *want* and *omdat* in Dutch.

This study investigates the idea of causal connectives as processing instructions. It is clear that these connectives indeed have an important role when it comes to the integration of discourse segments: They instruct the reader on how to connect the upcoming information to the information provided in the preceding clause. Moreover, our data suggest that these instructions are rather subtle in that they provide readers with precise information on the type of causal relation. We believe it is crucial for our understanding of the cognitive status of causality in discourse to account for the interaction between relations and their linguistic markers. Furthermore, it is imperative that we get a better grip on the exact time course of on-line processing. Eye-tracking results like the ones presented here contribute significantly to both these issues.

References

Baayen, R. H. (2008). Analyzing Linguistic Data. A Practical Introduction to Statistics Using R. Cambridge University Press.

[3] Of course it is not excluded that structural differences have some effect on processing, adding up to the difficulty due to difference in subjectivity.

BRITTON, B.K. (1994). Understanding Expository Text. Building Mental Structures to Induce Insights. In M.A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 641-674). San Diego, CA: Academic Press.

COUPER-KUHLEN, E. (1996) Intonation and clause-combining in discourse: The case of *because*. *Pragmatics* 6.3: 389-426.

COZIJN, R. (2000). *Integration and inference in understanding causal sentences*. Doctoral dissertation, Tilburg University.

DEGAND, L. (1996). *A situation-based approach to causation in Dutch with some implications for text generation.* Doctoral dissertation, Université Catholique de Louvain, Belgium

DEGAND, L., & SANDERS., T. (2002). The impact of relational markers on expository text comprehension in L1 and L2. *Reading and Writing, 15*, 739-757.

DEGAND, L., & PANDER MAAT, H. (2003). A contrastive study of Dutch and French causal connectives on the Speaker Involvement Scale. In A. Verhagen & J. van de Weijer (Eds.), *Usage based approaches to Dutch* (pp. 175-199). Utrecht: LOT.

KNOTT, A., & DALE, R. (1994). Using linguistic phenomena to motivate a set of coherence relations *Discourse Processes, 18*, 35-62.

KNOTT, A., SANDERS, T. . (1998). The classification of coherence relations and their linguistic markers: An exploration of two languages. *Journal of Pragmatics, 30*, 135-175.

MILLIS, K. K., & JUST, M. A. (1994). The influence of connectives on sentence comprehension. *Journal of Memory and Language, 33*, 128-147.

NOORDMAN, L. G. M., & DE BLIJZER, F. (2000). On the processing of causal relations. In E. Couper-Kuhlen & B. Kortmann (Eds.), *Cause - Condition - Concession - Contrast. Cognitive and discourse perspectives*. New York: Mouton de Gruyter.

NOORDMAN, L. G. M., & VONK, W. (1997). The different functions of a conjunction in constructing a representation of the discourse. In J. Costermans & M. Fayol (Eds.), *Processing interclausal relationships. Studies in the production and comprehension of text.* (pp. 75-93). Mahwah, New Jersey: Lawrence Erlbaum Associates.

PANDER MAAT, H., & DEGAND, L. (2001). Scaling causal relations and connectives in terms of speaker involvement. *Cognitive Linguistics, 12*(3), 211-245.

PANDER MAAT, H., & SANDERS, T. (2001). Subjectivity in causal connectives: An empirical study of language in use. *Cognitive Linguistics, 12*(3), 247-273.

PIT, M. (2003). *How to express yourself with a causal connective? Subjectivity and causal connectives in Dutch, German and French*. Amsterdam – New York: Editions Rodopi B.V.

Syntactic form at play with discourse relations

Bergljot Behrens, Kåre Solfjeld, Cathrine FabriciusHansen

Univ of Oslo, Østfold Univ. College, Univ of Oslo
{bergljbe,kare.solfjeld,c.f.hansen}@ilos.uio.no

Abstract Discourse relations have been defined for their contribution to text organization. The interplay between syntax and discourse relations has been little studied. By varying the syntactic form and relative position of propositions relating by discourse subordinating relations, we formulate contraints on competing realizations of the relation Background and Elaboration. Authentic data are extracted from the Oslo Multilingual Corpus.

Keywords: discourse relations, syntactic contratins, coordination, participles, punctuation

1 Introduction

In recent research increasing attention has been given to the interplay between syntactic form on clause/sentence level and discourse relations. In Tree Bank studies (Webber et al 2003) subordinate clauses have generally been correlated with discourse subordination, although closer empirical investigations in this area of text organization research indicates that the picture is more complex (Lee et al 2008). In SDRT, on the other hand, (Asher 1993, Asher, Prevot, Vieu 2008) discourse relations have normally been defined and analyzed on the basis of independent sentences, which may have either subordinate or coordinate discourse roles, depending to a large extent on the semantic relation that obtains between the eventualities they express. This does not mean that SDRT views independent sentences as the unique or only units relevant for discourse relational updating: coordinate compound clauses are taken up in the discussion as always rendering a discourse coordinate relation (Asher, Vieu 2005), and participle clauses as well as initial subordinate *while*-clauses and participle clauses are mentioned in a footnote as having the potential of relating to their matrix clauses by the discourse subordinate relation Background (Asher, Prevot, Vieu 2008).

Authentic, felicitous translation data between English, German, French and Norwegian demonstrate quite clearly that syntactic subordination competes with VP-coordination on the one hand, and independent juxtaposed sentences on the other. In the present paper we compare the discourse relation potential of intrasentential clauses as well as combinations of and with independent, juxtaposed sentences, and explore contraints that regulate/motivate their use.

If Asher and Vieu (2005) are correct in their analysis of syntactic coordination as always establishing a discourse coordinate role, how come a syntactically subordinate clause rendering a state and relating by discourse *subordinate* Background to its host, can be felicitously paraphrased by a finite first conjunct, but not by a second conjunct?

Furthermore, the discourse subordinate role of Elaboration may obtain between two propositions expressed syndetically or asyndetically. Elaborating adjuncts compete with elaborating independent sentences in translation. Is there any distinction between them?

The paper proceeds as follows: Before we go on to discuss the interplay between syntax and discourse organization in detail, we introduce the relevant discourse relations in section 2. Section 3 takes up syntactic constraints on inferring Background, with particular focus on their forward pointing direction. In Section 4 we take a closer look at syntactic constraints on backward pointing relations, and see them in a wider discourse context. Section 5 sums up our findings.

2 Relevant discourse relations

For precise definitions of the relevant relations we refer the reader to Asher (1993), Asher and Lascarides (2003), Asher et al (2008). Here we introduce them informally with their most central properties. Basically, different semantic relations between propositions, in principle irrespective of their syntactic realization, have different discourse roles: they are either coordinate or subordinate. Discourse subordination has the effect of allowing the proposition following the subordinate proposition to attach directly to, and thus create coherence with, the discourse dominating proposition. In a *Narrative* sequence, for example, governed by a temporal sequence of events, the propositions are discourse coordinate as long as they denote independent, temporally ordered events, while state descriptions, which may interfere with the narrative sequence, are discourse subordinate, yielding a *Background* relation to its surrounding event descriptions. Background descriptions can point forward, forming a Background for the following event narrated, or it may point backward, as a follow-up state description after an event has been introduced in the narrative (Asher et al 2008).

Discourse coordination also obtains in non-narrative discourse sequences, where a sequence of arguments (events or states) relate by *Continuation*, their sum potentially leading to a conclusion or *Consequence,* which is also discourse coordinate, closing off the continuation of arguments. Continuation as a discourse relation is thus similar to Narration, having the same characteristics but for the temporal aspect. Propositions relating by Continuation render lists of events or states-of affairs relevant in the same time space. However, if an event description is simply a specification of an eventuality presented in the preceding proposition, it relates to its preceding neighbor by *Elaboration.* The elaborating proposition is discourse subordinate to its more general superordinate description, allowing the next incoming event to potentially attach directly to the superordinate event.

3 Background in intrasentential clause combinations

Since discourse relations are defined for propositions irrespective of syntactic type, we are able to discuss the syntactic impact on text organization by varying the syntactic realization of the same sequences of propositions. In the following we consider constraints on the interpretation of Background in intrasentential clause combinations with coreferring subjects and the same tense. We understand Background in the sense of Asher et al. (2008), assuming that in narratives stative predicates function as either forward or backward pointing Background in relation to surrounding dynamic event predicates. We assume that unbounded events (activities/processes) can also function as Background - parallel to states.

Participial adjuncts are good candidates for Background, whether forward pointing as in (1) and (2), or backward pointing as in (3):

(1) *Von Liebe erfasst* schenkte der Jäger Atlante das ihm als Siegespreis zugesprochene Fell des Ebers.

(2) I nodded. He gave me a loaf of bread. ... When I woke up I found myself in a pit from which sand was excavated for the building of the road. I climbed out and fled through the forest.*Hugging what was left of my bread,* I went down the street.

(3) She stood watching Alice eating her salami and her pate on thick bread. Then watched while Alice peered into every corner of the refrigerator, and brought out some left-over spiced rice, which she ate with a spoon, *standing up.* (DL2)

The states expressed in the adjuncts above add circumstantial information surrounding the event expressed in the matrix clause.

An interesting observation is that VP-conjunction seems to allow forward-pointing Background (the Background state appearing in the first conjunct) (Behrens and Solfjeld, 2009) but blocks backward pointing Background. Thus an acceptable alternative to (1) above is (4) and the authentic translation of the adjunct in (3) into Norwegian restructures the information in a VP conjunction, placing the Background information in the first conjunct, as in (5):

(4) *Der Jäger war von Liebe erfasst* und schenkte…..

(5) *Hun sto rett opp og ned* og spiste grøten med skje
(She stood straight and ate the porridge with a spoon)

English too, seems to allow a forward pointing Background relation between conjuncts. (6) demonstrates a case of S-conjunction in which P2 relates to P3 by Background, according to the definition:

(6) Friday morning I got up at 6 and headed over to the beach for my run(P1). *It was cold* (P2) and I ran as much to keep warm as I did to keep in shape (P3).

VP conjunction has been claimed as the prototypical case of discourse coordination (Asher and Vieu 2005). In view of the German and Norwegian examples above we might consider language specific differences with respect to this constraint. However, the VP-conjoined clauses combining a state and a dymaic event in (7), slightly altered from (6), is a clear counterexample:

(7) Friday morning I got up at 6 and headed over to the beach for my run(P1). I was cold (P2) and ran as much to keep warm as I did to keep in shape (P3).

We have observed, then, that VP-conjunction seems to allow forward pointing Background. However, a VP-conjunction where a dynamic event is followed by a Background state in the second conjunct, is blocked. This does not mean that an e(vent) + s(tate) predicate combination in that order is ruled out in VP conjunction, viz.

(8) They all stopped eating *and sat very tense,* listening (RD1)

In English the state in the second conjunct either follows the event temporally, as in the example above, or it is re-interpreted as an event, as in (9):

(9) Mary came in *and was surprised.*

An attempted stative interpretation of the second conjunct makes the sequence pragmatically incoherent. Interpreted dynamically, however, the surprise can follow the entry temporally, and make sense. This automatic re-interpetation of the second predicate is interesting in that it differs from final participial adjuncts. (10), slightly changed from (9), with a participial adjunct, allows a backward-pointing Background reading:

(10) Mary came in *and was surprised.*

(10) is understood to mean that Mary was in a state of surprise when she came in. An authentic example is (11):

(11) But after that, everyone got up and moved off silently, *knowing that something important had just happened.*

The participial adjunct in (11) relates by Background to its host. Interestingly, in translation into German and Norwegian in which present participle clauses are highly constrained, this adjunct appears as an independent sentence:

(11’) Aber danach standen alle auf und zogen sich schweigend von der Veranda zurück, *alle wußten, daß etwas Wichtiges passiert war.*

(11’’) Men etter det reiste alle seg og gikk uten å si noe. *De var klar over at det hadde skjedd noe alvorlig.*

(12) [But afterwards everyone stood up and left without saying anything. They were aware that something serious had happened.]

Translation by VP conjunction would be infelicitous. This is confirmed by the questionable sequences (12’) and (12’’):

(12’) ? But after that, everyone got up, moved off silently and *knew that something important had just happened.*

(12’’) ? Aber danach standen alle auf, zogen sich schweigend von der Veranda zurück und *wußten, daß etwas Wichtiges passiert war..*

Hence, coordination is predicted not to occur in attested translations of final adjuncts encoding backward-pointing Background.

We can sum this up by placing the following syntactic constraint on Background:

> **Forward pointing Background can obtain intrasententially irrespective of clause type, while intrasentential backward pointing Background is restricted to syntactically downgraded structures, such as participial adjuncts.**

It might seem that the first part of the constraint is too wide, since it will allow cases in which the second clause is finite and subordinate. However, examples like (13) indicate that such cases should be included:

(13) The sun was low in the sky when the lady driver pulled out.

In (13) the state in the initial main clause forms the discourse background relative to the syntactically subordinate foreground event.

Initial adjuncts differ from their 1.conjunct counterparts, however, in also having the potential of establishing a discourse relation to its preceding context. While retaining its forward pointing Background relation to the following host, the initial adjunct *oben angekommen* '(having) arrived upstairs' in (14) expresses the already inferable result state of the actions narrated in the pre-context:

(14) Nach dem Gottesdienst ging ich auf die Empore, unter dem Vorwand, etwas mit dem Organisten besprechen zu wollen. Auf der Treppe begegneten mir die Mitglieder des Kirchenchors, aber ich sah keine, bei der ich mich gern dafür bedankt hätte, daß mein Sonntag von ihr vergoldet worden war.

Oben angekommen, begriff ich daß sie sich verspätet hatte, sie kehrte mir den Rücken zu und suchte ihre Sache zusammen.

(14') ?? ..... Ich kam *oben an* und begriff..../ ….I arrived upstairs and realized....

A finite coordinate alternative is not felicitous in such contexts, since events/states denoted by finite structures are explicitly asserted, and should therefore assert new information, not information already inferred from context. The non-finite initial adjunct has anaphoric properties that are exploited to create textual coherence between segments of text, like connectives. If the presupposition is not directly resolved in the preceding co-text, it will be accommodated.

4 Intrasentential or intersentential backward pointing relations.

We saw above that coordinate VPs are constrained to the effect that a stative predicate as second conjunct can not follow a dynamic event predicate in the first conjunct and at the same time yield a temporal overlap between the event and the state. So, locally, i.e. exclusively considering the two propositions standing in a Foreground-Background relation (in that order), only juxtaposed sentence sequences are felicitous competitors to final adjuncts with a Background role.

The discourse relation Elaboration, as understood by its very name, obtains only as a backward-pointing relation. It has generally been analyzed as a relation between propositions expressed in juxtaposed, independent sentences, in fact Asher and Vieu's analysis of VP coordination as discourse coordinate implies that Elaboration cannot obtain in coordinate compounds since it is discourse subordinate. In view of what we have seen above, this claim needs a slight modification: coordination is only blocked for backward-pointing discourse subordination. The claim has also been made that Elaboration can only obtain between independent sentences on grounds that the Elaborating proposition answers a question raised by the propositions elaborated on (Carston 2002). Such a claim is clearly invalidated as soon as non-finite clausal adjuncts are taken into account; as seen by for example (15):

(15) There were often six or seven of them all talking at once among the debris, *arguing about dates and availabilities* while Christian, the architect, acted as referee.

In German and Norwegian translations such Elaborations often appear in independent sentences, as expected viz.:

(16) Sie waren im ganzen sechs oder sieben und sprachen alle gleichzeitig inmitten der Trümmer, *sie diskutierten Termine und Verfügbarkeit,* während Christian, der Architekt, als Schiedsrichter fungierte.

(17) Det kunne være både seks og syv av dem samtidig, og alle snakket i munnen på hverandre i vårt kjøkkens ruiner. *De kranglet om leveringsdatoer og muligheter* mens arkitekten Christian fungerte som dommer.
[There might be both six and seven of them at the same time, and everybody spoke in each other's mouths in our kitchen's ruins. They argued about dates and possibilities while Christian acted as judge.]

So, independent sentences are adequate competitors to final adjuncts, and also the preferred finite alternatives under a Background or Elaboration reading. What, then, motivates the intrasentential form of final participial adjuncts?

The answer to this question is twofold: Backward-pointing Background seems more likely as adjuncts if the Background information load is low. Looking back to example (3) in section 3 above, we find the final adjunct as a short Background vomment on the subject's posture. Posture itself is implicitly given by the host, so very little information is actually added. This piece of added information would be odd in an independent sentence.

The case of Elaboration is more obscure: On the assumption that a full sentence exhausts the answer to the question under discussion, a truly elaborating adjunct should only be possible if the matrix host is felt by the speaker to be an incomplete answer to the question s/he wants it to answer. This is also indicated by the fact that Elaborating adjuncts often follow general posture verbs, such as sit, stand, lie, e.g:

(18) She stood in the doorway, *leaning slightly forward.*

(19) I lay in bed, *stretching out after a hard day's work.*

Or they are found with matrix predicates that are otherwise relatively void of descriptive content. Consider the following examples:

(20) [German tourists invading the country side of the French peasant] :
The German campers had treated his elaborate defence system with contempt, *rolling back boulders to make a gap in the barricade and stealing the notices that*

warned them of the presence of vipers (from Peter Mayle; A Year in Provence)

(21) [The truly French housewife at the market]:
She gets to grips with it, *snapping the match-thick thin haricots verts between her fingers, squeezing aubergines, sniffing tomatoes and tastig cheeses and olives.*

(22) She spent her days in relative but nonetheless real innocence, *shopping, rearranging her linen cupboard or her wardrobe, lunching with friends, strolling down Bond street.*

(23) Gold fever along the American River ran like a forest fire out of control, *bursting up one side of a ridge and down the other, spreading over regions north and south,* until waves of prospectors had scoured California's entire interior of foothills and flatland bounded by the Sierra Nevada to the east and the Coast Ranges to the west

The matrix clauses in the above examples are very general statements – in fact they all just name the subjectively evaluated type of event that the sum of the adjunct events represents. Such *criterial predicates* (Sæbø 2005, Behrens and Fabricius Hansen 2010) seem to favour adjunct specification, given a common subject. Notably, however, once the adjunct form has been chosen, a new independent sentence cannot continue the elaboration. Viz: the oddity of attempted paraphrases of (20) and (21) above:

(20’) ?The German campers had treated his elaborate defence system with contempt, *rolling back boulders to make a gap in the barricade. They stole the notices that warned them of the presence of vipers.*

(21’) ? She gets to grips with it, *snapping the match-thick thin haricots verts between her fingers. She squeezes aubergines, sniffs tomatoes and tastes cheeses and olives.*

A plausible interpretation of the empirical observations above is that of exhaustiveness (Jasinskaja 2004). The idea is that a full stop is a strong indicator that the information given in the sentence uttered is considered by the speaker to exhaust the answer to the question it raises, its QUD (Question Under Discussion) or Quaestio (Klein, Stutterheim 2001) i.e. the topic about which it gives new information. Jasinskaja (2004) uses this idea to account for the use of VP coordination: the two eventualities combined by conjunction together make up an exhaustive answer to the QUD. What we have seen here, however, is that it may account for participial adjuncts as well. Once the adjunct has been closed off with a full stop, the answer is exhausted. The information appearing in the sentence following the matrix-adjunct combination is therefore found incoherent if it cannot be understood as anything but a continuation of the elaborating adjunct The discourse coordinate Continuation relation, in other words, can only obtain with clauses of the same syntactic kind. This observation seems to us to be new.

We also observe that sequences of elaborating adjuncts always relate by Continuation, i.e. elaborating participial adjuncts are discourse coordinate amongst themselves, in the sense that they cannot elaborate on another of the same form. This becomes apparent by the rather infelicitous (24):

(24) ? Mary spent the whole day at home*, reading a book, delving into it with grea enthusiasm*

On the same grounds the following attempted adaptation of (20) above is infelicitous on a nested elaboration reading::

> (20’’) ?? The German campers had treated his defence system with contempt, *making a gap in the barricade, rolling back boulders.*

Finally, if discourse subordinate Elaboration holds between a participial adjunct and its matrix clause, full stop punctuation will block the elaborating proposition from being available as the sole attachment point for new incoming propositions.

> (20’’’) ??The German campers treated his defence system with contempt*, making a gap in the barricade. They rolled back boulders.*

This is explained on grounds of exhaustiveness: the information in the adjunct exhausts the elaboration. Since elaboration is already exhausted, signaled by the full stop, it does not make sense to continue with elaborating details.

Elaboration of an elaboration is consequently more successfully obtained in an independent sentence, as seen in the examples below::

> (25) The German campers treated the peasant’s defence system with contempt. They rolled back boulders, making a gap in the barricade.
>
> (26) The German campers treated the peasant’s defence system with contempt. They made a gap in the barricaede, rolling back the boulders.

Independent juxtaposed sententences, then, are preferred to adjuncts under a nested elaboration sequence.

5 Summing up our findings

Our data support the claim that there is an interplay between syntax and discourse relations, as different syntactic forms constrain the discourse relational potential.

Forward pointing Background can obtain intrasententially irrespective of clause type. VP-coordination and adjunction differ to the effect that state predicates in second VP-conjuncts following event predicates enforce a sequential interpretation, whereas states in final adjuncts are comapatible with temporal inclusion. This means that final adjuncts, but not second conjuncts, are available for backward-pointing Background. So, intrasentential backward pointing Background is restricted to syntactically downgraded structures, such as participial adjuncts. This means that a syntactically attached Background proposition points only in the direction of its intrasentential neighbor. Moreover, reduced initial structures have anaphoric properties that make them available as connectors, as opposed to first conjuncts,

Both independent sentences and final adjuncts may form Background/Elaboration relations to a preceding sentence/matrix clause. Low information load seems to favor adjunction to juxtaposition. Certain constraints have been observed, however: Sequences of final elaborating adjuncts can not be interpreted as elaborating each other, i.e. a series of final elaborating adjuncts are related by Continuation. Also, there is a constraint to the effect that an independent sentence can not enter into a Continuation relation with a preceding elaborating (final) adjunct, on account of the principle of exhaustification. Consequently, from an intrasentential perspective, further elaboration of the matrix - beyond one elaborating adjunct – must follow in the same form. Generally, this suggests a discourse principle to the effect that propositions related by Continuation must be of the same syntactic form.

So, choosing an adjunct to establish Background may be motivated by low information load, in initial position also by a wish to establish anaphoric relations and avoid asserting propositions already part of the common ground explicitly.. Choosing an adjunct in final position may motivated by a wish to establish what is felt by the speaker to be an exhaustive answer to the Quaestio Once an adjunct is chosen for Elaboration, the same form must be used for Elaboration to be Continued. Conversely, making a full stop signals that the sentence makes up an exhaustive answer not in need of further elaboration. A new incoming proposition will relate to the preceding complex proposition, not to any of its constituent parts.

References

ASHER N (1993) *Reference to Abstract Objects* Dodrecht Kluwer

ASHER N , LASCARIDES A (2003) *Logics in Conversation* New York Cambrifge University Press

ASHER N, VIEU L.(2005) Subordinating and Coordinating Discourse Relations *Lingua* 115, 591 -610

ASHER N, PREVOT L, VIEU L.(2008) Setting the Background in Discourse. *Discours* 1, Toulouse

BEHRENS B, FABRICIUS-HANSEN C (2010) The relation Accompanying Circumstance across Languages, in SHU D, TURNER K eds *Contrasting Meaning in Languages of the East and West* Oxford, Berlin Peter Lang 531 - 551

BEHRENS B, SOLFJELD K (2009) Discourse role guiding translation choice, Conference proceedings, *LTSP* Paris

CARSTON R (2002) *Thoughts and Utterances The Pragmatics of Explicit Communication* Oxford Blackwell

JASINSKAJA K (2004) Exhaustification and Semantic Relations in Discourse in GEURTS , VAN DER SANDT eds. *Proceedings on Workshop on Implicature and Conversational Meaning* 14-19

LEE A, PRASAD R, JOSHI A, WEBBER B (2008) Departures from Tree Structures in Discourse – in www. *seas.upenn.edu/~pdtb/papers/**lee**-etal08-cid3.pdf*

KLEIN W, VON STUTTERHEIM C, (1991) Text Structure and Referential Movement, in *Sprache und Pragmatik 22* 1-32

SÆBØ K J (2007) The Structure of Criterion Predicates in DÖLLING J, ZYBATOV G (eds) *Event Structures in Linguistic Form and Interpretation* Berlin, Walter de Gruyter

VON STUTTERHEIM, C (1997) *Einige Prinzipien des Textaufbaus.* Tübingen: Niemeyer

WEBBER B, STONE, B JOSHI A, KNOTT A (2003) Anaphora and Discourse Structure, *Computational Linguistics* 29(4) 545-587

Cohesive Links with Literal and Idiomatic Expressions in Discourse: An Empirical and Computational Study

Caroline Sporleder, Linlin Li, and Alexis Palmer
Saarland University, Postfach 15 11 50
66041 Saarbrücken, Germany
{csporled, linlin, apalmer}@coli.uni-saarland.de

Abstract. Lexical cohesion is an important device for signalling text organisation. In this paper, we investigate to what extent idiomatic expressions participate in the cohesive structure of a text. We look at the problem both from an empirical and a computational perspective. Our results show that both literal and nonliteral expressions exhibit cohesion with their context, though the latter tend to do so to a lesser extent. We also show that cohesive links identified by humans can be approximated by an automatically computable semantic relatedness measure based on search engine page counts.

Keywords. idioms, multi-word expressions, cohesion, semantic relatedness, lexical chains, natural language processing, annotation

1 Introduction

The term ‘cohesion’ refers to the manner in which words or syntactic features connect individual sentences and clauses to their discourse context. Halliday and Hasan (1976) propose five classes of cohesion: conjunction, reference, substitution, ellipsis, and lexical cohesion. In this paper we focus on lexical cohesion, which covers various kinds of semantic relationships between the lexical items in a text. These range from literal repetition (called ’reiteration’ by Halliday and Hasan) to weaker semantic relationships (so-called ‘non-classical relations’ (Morris and Hirst2004)), as between *wet* and *dry* or *laugh* and *joke*.

Lexical cohesion is important because it is typically the most frequent class of cohesive ties (Hoey1991). It is also interesting because it tends to be global, i.e., texts can be analysed in terms of *chains* of lexically cohesive words. These chains can span large segments of the text or even the text as a whole, if the chain refers to the central topic of a discourse. Lexical ties thus indicate the overall structure and organisation of a text, e.g., in terms of the main topics addressed and their distribution throughout the text. Lexical cohesion has received considerable attention from both the linguistics (Hoey1991; Tanskanen2006) and the computational linguistics communities. Computational applications that make use of lexical chains range from the detection of malapropisms (Hirst and St-Onge1998) over word sense disambiguation (Okomura and Honda1994) and topic segmentation (Hearst1997) to automatic text summarisation (Barzilay and Elhadad1997).

In this paper, we address a particular aspect of lexical cohesion, namely how idiomatic expressions fit into the cohesive structure of a text. We are interested in (i) whether it is possible to find cohesive ties between a figuratively-used expression such as *break the ice* and the surrounding context, (ii) whether such ties are stronger or weaker than for the component words of the literal counterpart of the expression, and (iii) how such cohesive links can be modelled computationally. The latter point is important because recent work by Sporleder and Li (2009) suggests that the cohesive structure of a text can provide cues for the detection of non-literal language. We evaluate whether the cohesive links found automatically are identical or at least similar to those annotated by humans. We explore whether deviations between the two are due to errors made by the automatic method or whether humans pick up on a different type of cohesion than is captured by the automatic tool. Our work thus combines empirical and computational approaches.

2 Human evaluation of cohesive chains

To identify cohesive links, we carried out a small-scale annotation study, using texts from Sporleder and Li's (2009) dataset. We chose five expressions which can be used literally as well as idiomatically: *bounce off the wall* (henceforth: **wall**), *get one's feet wet* (**feet**), *rock the boat* (**boat**), *break the ice* (**ice**), and *play with fire* (**fire**). The expressions were chosen for different reasons: for *bounce off the wall*, Sporleder and Li's method erroneously classified many literally used examples as non-literal, and the reverse is true for *get one's feet wet*. *Rock the boat* was included because the performance of the classifier was relatively high for this expression; *break the ice* and *play with fire* were selected more or less randomly. For each expression, we randomly chose four texts from the corpus for annotation: two with literal uses and two with non-literal uses. Two annotators labeled the complete texts, and a third annotator labeled the portion of text immediately surrounding the expression of interest, with a window of approximately two paragraphs in each direction.

To test the hypothesis that literal and non-literal meanings of an expression can be distinguished on the basis of lexical chains, we annotated two chains for each text: one for the literal meaning of the target expression (**literal chain**) and one for the idiomatic meaning (**idiomatic chain**). Our expectation is that one chain—the one for the meaning intended by the author—should always be noticeably stronger than the other; cohesion with the non-intended meaning should be merely accidental and one might expect that authors try to deliberately minimise it to avoid confusion.[1] Our second hypothesis is that idiomatic usages tend to exhibit weaker cohesion (with the idiomatic chain) than literal usages do (with the literal chain).

2.1 Annotation process and decisions

Annotating cohesive chains is a notoriously difficult task, since it is often a matter of debate whether, to what degree, and in what way two words are semantically related. Relatively few empirical studies have looked into human intuitions regarding lexical cohesion, and those that have generally report a relatively low inter-annotator agreement (Morris and Hirst2005; Hollingsworth and Teufel2005; Klebanov and Shamir2006). To alleviate this problem to an extent, the annotators in our study discussed some of the early texts post-annotation, with the

[1] Sometimes one can observe intended cohesion with both meanings, usually due to a deliberate play with words.

1ST ANCHOR		2ND ANCHOR	
bounce	**15**	wall	6
rock	5	boat	**70**
break	13	ice	**36**
feet	20	wet	**89**
play	3	fire	**38**

Table 1: **Literal chains:** Number of cohesive links to anchor words of target expressions

aim of detecting potential problems and arriving at some general guidelines for the task. For this initial study, we include those texts in the results discussed below.[2]

After some discussion, the decision was made to mark shallow, lexically-based semantic relationships between words. Cohesive links based on world knowledge (for example, linking *pasta* with *marathon* via knowledge of the practice of carb-loading) were not marked, and named entities were also left unmarked. In addition, only content words were considered eligible for participation in cohesive chains.

Literal chains. Annotators identified **literal** and **idiomatic** chains for all texts, regardless of whether the target expression itself is used with its literal or its idiomatic meaning. Two anchor words were identified for each idiom, corresponding to the semantically most contentful words of the expression, e.g., a verb and a noun in V+NP or V+PP constructions. Annotators marked literal cohesion chains for both anchor words. The anchor words and the number of links to each appear in Table 1. In all but one case, the second anchor word, typically a noun, receives many more cohesive links than the first. This confirms an intuition that nouns exhibit more cohesion with their context or at least participate in more easily identifiable cohesion relations than verbs.

Idiomatic chains. For the idiomatic chains, annotators marked words exhibiting lexical cohesion with the idiomatic meaning of the target expression. Because that meaning can be difficult to pin down, we developed a set of paraphrases for each idiom. These paraphrases were used both to guide human annotation and for automatic computation of cohesion (see Section 3, Table 3).

Gold standard. A gold standard set of annotations was produced by adjudication over the individual annotations. We distinguished two types of cohesive links: weak and strong. Strong links were annotated for strong semantic relationships, such as that between *wet* and *water*. Weak links were annotated for more indirect relationships, e.g., between *wet* and *diving*, which are related via the concept of *water*.

Multi-word expressions The human annotators marked relevant multi-word expressions (MWEs) as participating in cohesive links, with each expression representing a single link. However, MWEs pose a particular challenge for automated text-processing systems, and the method we use to compute lexical cohesion does not accommodate MWEs. This has the result that some prominent cohesive links will be ignored in the automated processing. (1) is from

[2] Table 2 shows the respective annotator agreement levels for discussed and undiscussed texts.

Texts	Literal chains	Nonliteral chains
ALL	0.8115	0.7354
Literal usage	0.8189	0.6724
Nonliteral usage	0.8031	0.7747
Lit.discussed	0.8142	0.6639
Lit.not.disc	0.8235	0.7061
Nonlit.discussed	0.9594	0.9388
Nonlit.not.disc	0.6859	0.6763

Table 2: Correlations between annotators 1&2

a text with an idiomatic occurrence of *get one's feet wet*. The text is a report on the small but growing number of women in talk radio and the obstacles they face on that career path.

(1) That's not due to gender bias, although **breaking into the field** is harder for a woman, McCoy said. 'I think it might be tougher for a woman to **get started** than a man.'

Both *break into the field* and *get started* are reasonable (though not perfect) paraphrases for the idiom and as such form strong cohesive links. Of the individual words in the two phrases, only one of each link (*field* and *started*) independently exhibits lexical cohesion with the idiomatic meaning.

In other cases, though, each content word of a MWE exhibits cohesion with the target expression. In those cases, the links are preserved by marking each word separately. (2), from the same text as (1), shows two such cases (*enter the field* and *developing skills*).

(2) That is changing, though, as more women **enter** the **field**... Now that more are, they are *getting their feet wet* and **developing skills**.

2.2 Findings

To determine the reliability of our annotation, we computed the correlation between the first two annotators using Pearson's product-moment correlation, as implemented in R. The top half of Table 2 shows aggregate correlation figures for all texts, broken down between those with literal uses of the target expressions and those with non-literal uses. The bottom half of the same table distinguishes texts which were discussed by the annotators from those which were annotated entirely independently.

It can be seen that the correlation is generally relatively good, even for the texts that were not discussed. Overall, the correlation is higher for literal than for non-literal chains. Hence it seems that is easier to agree on related words for literal usages, while non-literal usages are fuzzier and therefore less easy to annotate.

Once the human annotations had been adjudicated and a gold standard produced, we computed the strengths of the annotated chains. It is common in computational linguistics to model the strength of a chain in terms of its length, i.e., the more word tokens a chain contains, the stronger it is. We adopted this measure here.[3] As expected, the chains for the intended meaning tend

[3] It should be noted that this measure does not make any distinction between the strength of individual links. In other words, weak and strong links count equally when determining strength of the chain.

to be stronger than those for the non-intended meaning, and this is true for both literal and non-literal usages.

Of the ten idiomatic usages, eight have stronger non-literal chains than literal chains. The first exception is a text about a diver who is "getting his feet wet" in the diving profession (see Example (4) below). Here the idiom is clearly used tongue in cheek, and the cohesion with the literal meaning is probably intentional. In the second exceptional text, the strong cohesion with the literal reading is probably accidental. The text contains a non-literal usage of "playing with fire," and the main topic of the text deals with bombs and rockets, which both annotators marked as being weakly related to *fire*.

Of the ten texts with literal usages, nine have stronger literal chains than non-literal chains. The single exception is a **bounce** text about car racing, in which the annotators found weak links between the non-literal meaning and words like *boring*, *slow*, and *speed*.

The results also confirm our second hypothesis: that idiomatic usages generally tend to exhibit lower degrees of cohesion with the text containing them than do literal usages. However, for most idiomatic usages, the annotators marked some words in the context as being related to the non-literal meaning. Hence, even idiomatic usages participate in cohesive relations with the context. At the same time, there tend to be fewer of these than for literal usages, and the relations tend to be weaker and more indirect (as indicated by the lower inter-annotator agreement seen for non-literal chains, Table 2).

All in all the results of our annotation study confirm the hypothesis that literal and non-literal usages can be distinguished based on the cohesive relationships they enter into with their texts: strong literal chains indicate literal usages, and if the idiomatic chain is stronger than the literal one, it is more likely that the expression is being used idiomatically.

2.3 Mixed literal and non-literal use

Most of the time, determining whether one of the target expressions is being used literally or idiomatically is a straightforward task. However, we encountered several interesting cases which seem to combine literal and non-literal uses. In these cases, it is more difficult to pull apart the interactions between the two cohesion chains. Here we discuss two examples.

Metaphorical 'literal' expressions. The passage in (3) is taken from one of the **wall** texts. In this case, the phrase is used in its literal sense, but situated in a rich metaphorical context.

(3) That movie was entertaining in an off the wall way. "If Lucy Fell" **bounces off the wall** and drops to the floor like a pound of old fish.

The first sentence of this passage uses the idiom *off the wall*, which may be related to the target expression *bounce off the wall* but clearly has a distinct meaning. This is then echoed (via repetition of the last three words) in the second sentence, where the target expression occurs in a pseudo-literal usage. We call this 'pseudo-literal' because it is meant to evoke the image of something wet and smelly hitting a wall and sliding down it. In this case, though, it is the movie which is (metaphorically) said to be sliding down the wall.

Signalling mixed use. Other interesting cases arise when the writer selects an idiom whose literal meaning relates to the topic of the text. One of the **feet** texts is about a man changing careers from drifter to diver. The target expression is used idiomatically in reference to one of his early diving jobs. This is one of the exceptions to the general rule that the chain for the primary intended meaning should be stronger. The literal chain here contains 21 tokens, while the non-literal chain has only six.

(4) Davila said he worked for Disney to **get his feet wet**, so to speak.

In this case the phrase *so to speak* is used to draw attention to the nature of the use of the idiom, which almost has the feeling of a pun, due to its semantic proximity to the topic of the article. The phrase is one way of signalling a 'complex' usage of the idiom, where the main meaning is non-literal but there is also strong lexical cohesion between the text and the literal meaning of the expression.

In future work it would be interesting to explore the role and distribution of *so to speak* and similar cue phrases (e.g. *if you will*, *in a manner of speaking*, *as it were*). On cursory examination, such phrases often occur as a rhetorical strategy to express the author's awareness of the potential for multiple interpretations of the expression of interest, and perhaps also to call the reader's attention to that potential. One interesting question is whether they may also serve to point toward the topic of the text, suggesting that the literal meaning is a prominent theme in the text.

3 Automatic Methods

Our results from the human annotation study suggest that literal and idiomatic usages of an expression can indeed be distinguished on the basis of their cohesive links with the surrounding text. However, this is only useful for automatic idiom detection if such cohesive links can be identified automatically. To be able to do so requires a measure of semantic relatedness that can be computed automatically for pairs of words. Modelling semantic relatedness is a very active research area in computational linguistics and various relatedness measures have been proposed and used in previous research. We chose a measure called *Normalized Google Distance* (NGD, see (Cilibrasi and Vitanyi2007)), since it has been used in an idiom detection task before (Sporleder and Li2009) and has the advantage of not being restricted to classical relations. NGD computes relatedness on the basis of page counts returned by an internet search engine. The basic idea is that the more often two terms occur together relative to their overall occurrence frequency the more closely related they are. NGD is defined as follows:

$$NGD(x, y) = \frac{max\{log\ f(x), log\ f(y)\} - log\ f(x, y)}{log\ M - min\{log\ f(x), log\ f(y)\}} \tag{5}$$

where x and y are the two words whose association strength is computed (e.g., *fire* and *coal*), $f(x)$ is the page count returned by the search engine for the term x (and likewise for $f(y)$ and y), $f(x, y)$ is the page count returned when querying for $x\ AND\ y$ (i.e., the number of pages that contain both, x and y), and M is the number of web pages indexed by the search engine. Note that NGD is a measure of *distance*, i.e., a low value means that two words are rated to be very similar.

Cohesion with literal reading. To model the text's lexical cohesion with the literal usage, we computed the NGD between the content words of each target expression and all other content words in the text. A known drawback of measures like NGD is that search engines do not always produce reliable page counts for high-frequency words (see Sporleder and Li (2009)). For this reason, we were not able to compute the cohesive structure for *play with fire*, as the search engine had trouble with both anchor words.

Cohesion with nonliteral reading. Sporleder and Li (2009) computed cohesion with only the literal chain and then predicted literal usage if this chain was strong and idiomatic usage otherwise. For this study, we are interested in also computing cohesion with the nonliteral chain.

The meaning of a literally used expression is compositional and thus relatively easy to model. The idiomatic meaning is more difficult to model since an explicit semantic representation is missing. In our experiment, we compared two methods for computing cohesive links for nonliteral meaning: (i) by using the full string of the target expression, and (ii) by using human-generated paraphrases of the idiomatic meanings.

The motivation for using the full string of the target expression is based on a study by Riehemann (2001), who found that expressions in canonical form (i.e., the dictionary form of an idiom) are more likely to be used idiomatically than literally. Hence, while the pages returned by querying for the full string of the target expression (i.e., the canonical form) may contain some literal usages, the majority of pages should contain idiomatic usages.

Querying for the full string gets relatively low page counts since the frequency of the full expression is usually much lower than that of its parts. We also found that idiomatic readings tend to appear in rather diverse contexts. For instance, *rock the boat* can mean *cause trouble* or *go against conventions*. It is more likely that words such as *accusation, attack, conflict* co-occur with the first reading, while words such as *counterculture, rebels, change, norm* co-occur with second reading. The diversity of nuances to the idiomatic meaning leads to a scattered distribution of the idiomatic meaning across many different context words. As a result, the nonliteral NGD is generally high (i.e., words tend to be rated as not very similar to the idiomatic meaning). This actually closely resembles human intuition, in that humans also rate cohesive links with idiomatic meanings as relatively weak.

In addition to using the full string to model idiomatic meaning, we also employed human-generated paraphrases, which were then used instead of the full string when querying the search engine. Intuitively, this method should lead to better results as paraphrases make it possible to pinpoint the meaning of an idiom more precisely. Table 3 shows the paraphrases we used. We deliberately tried to use short expressions in the paraphrases. Computationally, we represent the idiomatic reading by using the **OR** logic operator to connect all the possible paraphrases when sending a query to the search engine.

Comparing the results obtained by using the full-string model to those of the paraphrase model, we found evidence that the latter is more suited to modelling idiomatic meaning. Using paraphrases generally leads to lower NGD values, i.e., more words from the text are rated as being semantically related to the idiomatic meaning. Furthermore the words rated as similar to the target meaning seemed more plausible than those returned by the full-string model. We thus used the paraphrase model in our final experiment described below.

Idiom	Paraphrases
bounce off the wall	“high-strung”, “energetic”, “over excited”
get one’s feet wet	“first experience”, “dabble”, “dabbling”
rock the boat	“upset conventions”, “break norms”, “cause trouble”, “disturb balance”
break the ice	“ease tensions”, “get people talking”, “facilitate communication”
play with fire	“risky behaviour”, “risky behavior”, “take risks”, “act dangerously”

Table 3: Paraphrases for idiomatic meanings

3.1 Manually vs. Automatically Found Cohesive Links

In our final experiment, we compared the cohesive links in the manually created gold standard to those found automatically by the method described above. Figures 1 to 4 plot the NGD for a given word against its position in the text. This allows us to see whether there are more and stronger cohesive links with words in the local vicinity of the target expression. The position of the target expression in the text is marked by a (blue) vertical line. Words that were marked as semantically related in the gold standard are indicated by a (green) bullet. Figure 1 show the results for the literal chain of a literal usage of *rock the boat*, while Figure 2 shows the results for the nonliteral chain for the same literal usage of *rock the boat*. Similarly, Figures 3 and 4 show the chains for an idiomatic usage of *rock the boat*; the former depicts the nonliteral chain, i.e., the chain for the intended usage, the latter shows the literal chain.

The first observation that can be made is that the position of a word in the text relative to the target expression does not seem to correlate with its likelihood to form a cohesive link, i.e., the related words tend to be scattered all over the text and do not just appear in the neighbourhood of the target expression. This is true both for the human annotation (i.e., there are several links with words far away from the target expression), and for the automatically computed NGD (i.e., the NGD is not necessarily lower in the vicinity of the target).

Second, it can be seen that human annotations agree quite well with the NGD values; words marked by humans tend to be located at local minima in the graph. Humans thus often mark those words whose NDG is relatively small, i.e., which are rated as semantically similar to the target expression. This general pattern is observable for both the idiomatic and the literal meaning. Hence, it seems that modelling idiomatic meaning by combining NGD with human-generated paraphrases is a good strategy.

The results confirm that the literal cohesion is stronger than the nonliteral cohesion. While most of the chain words in the literal chain have an NGD value below 0.5, most of the words in the nonliteral chain have an NGD value around 0.8. This is also in line with our findings for the human study.

the literal chain than the nonliteral chain (see Figure 1 vs. 2 and Figure 3 vs. 4). We also find that if there are only a few words in the nonliteral chain, they are more likely to appear in the neighbourhood of the target expression (see Figure 4). This means that the nonliteral reading can be obtained within a relatively small context.

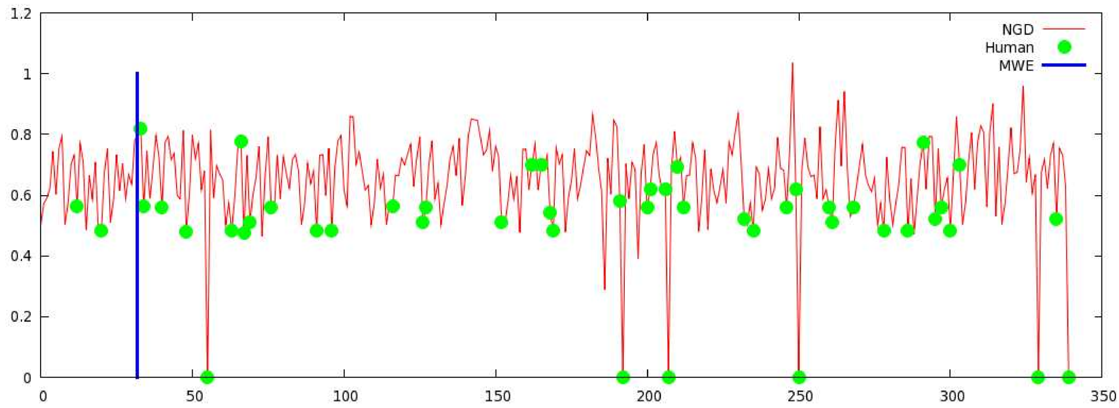


Figure 1: Example of a literal chain for a literal usage ("rock the boat"). The **x** axis represents the position of the tokens in the text. The **y** axis is the NGD value between the token and the literal reading of the target expression (MWE)

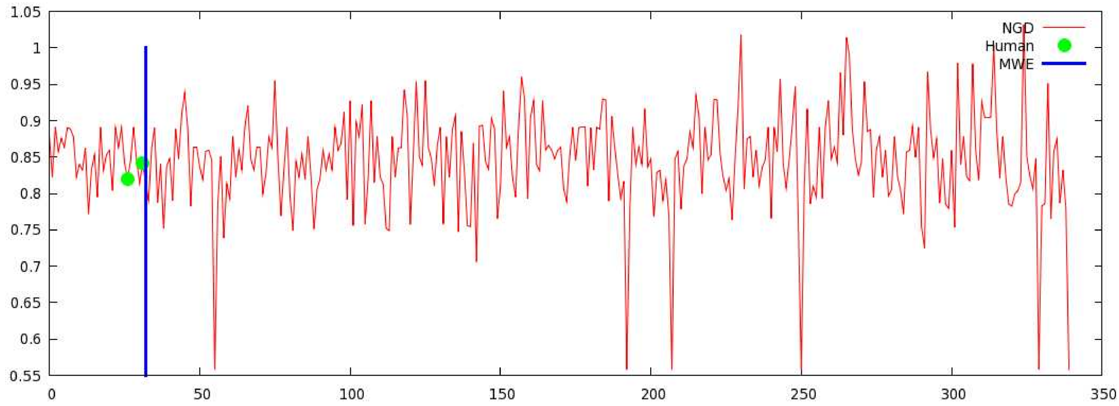


Figure 2: Example of a nonliteral chain for the same literal usage as Figure 1 ("rock the boat"). The **x** axis represents the position of the tokens in the text. The **y** axis is the NGD value between the token and the nonliteral reading of the target expression (MWE)

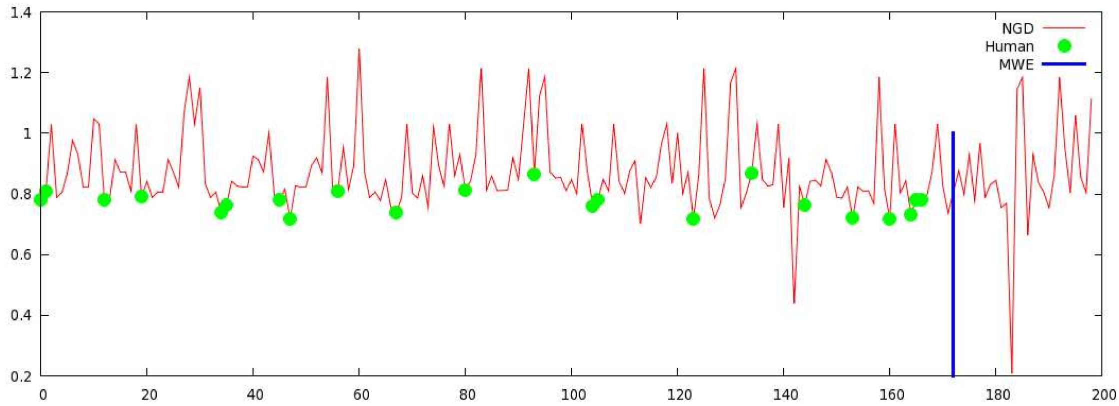


Figure 3: Example of a nonliteral chain for a nonliteral usage ("rock the boat"). The **x** axis represents the position of the tokens in the text. The **y** axis is the NGD value between the token and the idiomatic reading of the target expression (MWE)

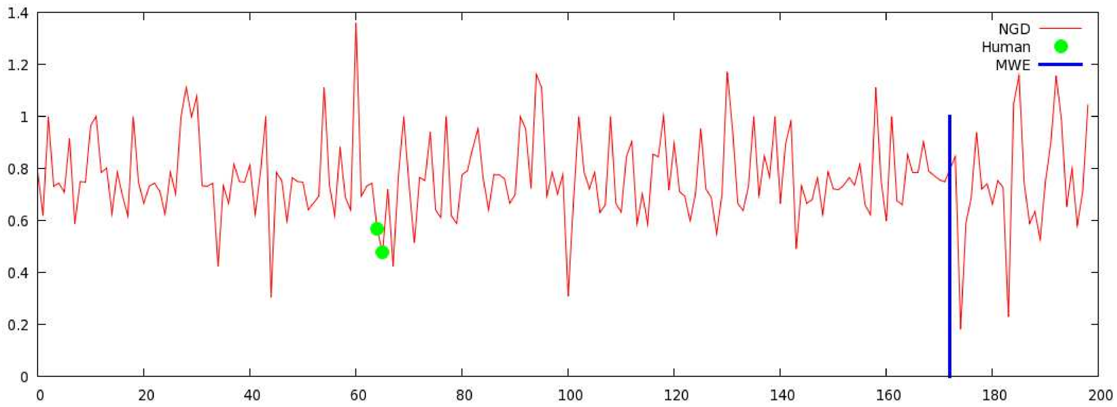


Figure 4: Example of a literal chain for the same nonliteral usage as Figure 3 ("rock the boat"). The **x** axis represents the position of the tokens in the text. The **y** axis is the NGD value between the token and the literal reading of the target expression (MWE)

4 Conclusions

In this study, we addressed the question of how idiomatic and literal meanings participate in the cohesive structure of a text. Our findings suggest that both literal and non-literal meanings exhibit lexical cohesion with their context, however for non-literal meanings the cohesive ties tend to be much weaker. Links with the non-intended reading of an expression are typically weak, hence the cohesive structure of a text can be used to distinguish literal and non-literal readings. One exception arises in cases where an idiom is used tongue in cheek, i.e., it is deliberately chosen to cohere with both meanings.

We also investigated whether cohesive chains can be computed automatically. We found that a distance measure based on internet search engine page counts produces good results. Furthermore, it seems that with this method the non-literal meaning of an expression can be modelled well by human-generated paraphrases.

In ongoing work, we are annotating a larger data set to explore the cohesive links in texts more fully. We are particularly interested in those cases where a deliberate play with words on the part of an author means that an expression exhibits cohesive links under both the literal and non-literal reading.

Acknowledgments

This work was funded by the German Research Foundation DFG within the Cluster of Excellence "Multimodal Computing and Interaction" (MMCI).

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for summarization. In *Proceedings of the ACL-97 Intelligent Scalable Text Summarization Workshop (ISTS'97)*.

Rudi L. Cilibrasi and Paul M.B. Vitanyi. 2007. The Google similarity distance. *IEEE Trans. Knowledge and Data Engineering*, 19(3):370–383.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Marti Hearst. 1997. Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database*, pages 305–332. The MIT Press.

Michael Hoey. 1991. *Patterns of Lexis in Text*. Oxford University Press, Oxford.

Bill Hollingsworth and Simone Teufel. 2005. Human annotation of lexical chains: Coverage and agreement measures. In *Proceedings of the SIGIR-05 Workshop ELECTRA: Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications*.

Beate Beigman Klebanov and Eli Shamir. 2006. Reader-based exploration of lexical cohesion. *Language Resources and Evaluation*, 40(2):109–126.

Jane Morris and Graeme Hirst. 2004. Non-classical lexical semantic relations. In *Proceedings of the HLT-NAACL-04 Workshop on Computational Lexical Semantics*, pages 46–51.

Jane Morris and Graeme Hirst. 2005. The subjectivity of lexical cohesion in text. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing attitude and affect in text*. Springer.

Manabu Okomura and Takeo Honda. 1994. Word sense disambiguation and text segmentation based on lexical cohesion. In *Proceedings of the 15th International Conference on Computational Linguistics (Coling-94)*.

Susanne Riehemann. 2001. *A Constructional Approach to Idioms and Word Formation*. Ph.D. thesis, Stanford University.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)*.

Sanna-kaisa Tanskanen. 2006. *Collaborating Towards Coherence: Lexical Cohesion in English Discourse*. John Benjamins.

Index

MAD 2010 - List of Participants

ADAM Clémentine
PhD Student
University of Toulouse-Le Mirail, France
clementine.adam@univ-tlse2.fr

AMADIEU Franck
Senior Lecturer in Psychology
University of Toulouse-Le Mirail, France
amadieu@univ-tlse2.fr

BATEMAN John
Professor of Applied English Linguistics
University of Bremen, Germany
bateman@uni-bremen.de

BEHRENS, Bergljot
Associate Professor
University of Oslo, Norway
bergljbe@ilos.uio.no

CANESTRELLI Anneloes
PhD Student
Utrecht Institute of Linguistics OTS,
Netherlands
a.r.canestrelli@uu.nl

DANLOS Laurence
Professor of Computational Linguistics
Paris Diderot University, France
laurence.danlos@linguist.jussieu.fr

DELIN Judy
Professor of Language & Communication
University of Reading, UK
judy.delin@roedelin.com

DE WAARD Anita
Disruptive Technologies Director
Utrecht Institute of Linguistics OTS,
Netherlands
a.dewaard@elsevier.com

FABRE Cécile
Senior Lecturer
University of Toulouse-Le Mirail, France
cecile.fabre@univ-tlse2.fr

FABRICIUS-HANSEN Cathrine
Professor
University of Oslo, Norway
c.f.hansen@ilos.uio.no

GIRAUD Stéphanie
PhD Student, Ergonomist
University of Nice - Sophia Antipolis, France
stephanie@ludo-tic.com

HERMANN Ėric
Associate Director
Conseils en Facteurs Humains (CFH), France
hermann.cfh@orange.fr

HO DAC Lydia-Mai
Post-doctoral Researcher
Catholic University of Louvain, Belgium
hodaclm@gmail.com

LAIPPALA Veronika
PhD Student
University of Turku, Finland
veronika.laippala@utu.fi

LAVID Julia
Professor in English Linguistics
Universidad Complutense de Madrid, Spain
lavid@filol.ucm.es

LE DRAOULEC Anne
Research Scientist
University of Toulouse-Le Mirail, France
draoulec@univ-tlse2.fr

LEMARIĖ Julie
Senior Lecturer in Psychology
University of Toulouse-Le Mirail, France
lemarie@univ-tlse2.fr

LONGRĖE Dominique
Professor
University of Liège, Belgium
dominique.longree@ulg.ac.be

LORCH Robert F.
Professor of Psychology
University of Kentucky, USA
rlorch@email.uky.edu

MACKSOUD Ruby
Instructor
Arizona State University, USA
ruby.macksoud@asu.edu

MATHIAS Peter
Managing Director
Bridge Research and development, UK
petercmathias@btinternet.com

MAUREL Fabrice
Senior Lecturer
University of Caen, France
fmaurel@info.unicaen.fr

MELLET Sylvie
Senior research fellow CNRS
University of Nice - Sophia Antipolis, France
mellet@unice.fr

OAKEY David
Assistant Professor of Applied Linguistics
Iowa State University, USA
djoakey@iastate.edu

PALMER Alexis
Post-doctoral Researcher
Saarland University, Germany
apalmer@coli.uni-sb.de

PĖRY-WOODLEY Marie-Paule
Professor of Linguistics
University of Toulouse-Le Mirail, France
pery@univ-tlse2.fr

PIMM Christophe
Linguist
Conseils en Facteurs Humains (CFH), France
cpimm@univ-tlse2.fr

REBEYROLLE Josette
Senior Lecturer
University of Toulouse-Le Mirail, France
rebeyrol@univ-tlse2.fr

ROZE Charlotte
PhD Student
Paris Diderot University, France
charlotte.roze@linguist.jussieu.fr

SAINT-DIZIER Patrick
Senior research fellow CNRS
Institut de Recherche en Informatique de Toulouse, France
stdizier@irit.fr

SARDA Laure
Research Scientist, CNRS
LATTICE, France
laure.sarda@ens.fr

SPORLEDER Caroline
Research group Leader
Saarland University, Germany
csporled@coli.uni-saarland.de

THĖROUANNE Pierre
Senior Lecturer
Université de Nice - Sophia Antipolis, France
therouan@unice.fr

THOMAS Martin
Research Assistant
University of Reading, UK
m.thomas@leeds.ac.uk

VANDI Claudio
PhD Student
Paris 8 University, France
vandi@lutn-userlab.fr

VERGEZ-COURET Marianne
PhD Student
University of Toulouse-Le Mirail, France
Marianne.Vergez@univ-tlse2.fr

WALLER Robert
Professor of Information Design
University of Reading, UK
r.waller@reading.ac.uk

ZAFIHARIMALALA Herimanana
PhD Student
University of Toulouse-Le Mirail, France
Herimanana.Zafiharimalala@univ-tlse2.fr

3 NoteBook

MAD 2010 - PROGRAMME

WEDNESDAY, MARCH 17	THURSDAY, MARCH 18	FRIDAY, MARCH 19	SATURDAY, MARCH 20
	9:15	**9:00**	**9:15**
	Waller, R. & Delin, J. Towards a pattern language approach to document description **Vandi, C. & Baccino, T.** Spatial Coding and information retrieval in multimodal documents	**Roze, C., Danlos, L. & Muller, P.** LEXCONN: A French Lexicon of Discourse Connectives **Fontan, L. & Saint-Dizier, P.** Text Organization: Identifying and Measuring the Strength of Arguments in Procedural Texts	**Canestrelli, A., Sanders, T. & Mak, P.** Comparing *because* to *want*; How connectives affect the processing of causal relations. **Behrens, B., Solfjeld, K. & Fabricius-Hansen C.** Syntactic form at play with discourse relations
	Coffee Break: 10:35-10:50	***Coffee Break: 10:20-10:35***	***Coffee Break: 10:35-10:50***
	Zafiharimalala, H. & Tricot, A. Text signals in the aircraft maintenance documentation **11:30:** ***Éric Hermann*** *NLP tools applied to Accident Incident Reporting Systems*	**de Waard, A.** Realm Traversal In Biological Discourse: From Model To Experiment and ... **11:15:** ***Robert F. Lorch*** Effects on Text Processing of Signaling Text Organization	**Sporleder, C., Li, L. & Palmer, A.** Cohesive Links with Literal and Idiomatic Expressions in Discourse: An Empirical and Computational Study **11:30:** ***Summing up & Closing***
	LUNCH: 12:3O-2:00pm	***LUNCH: 12:15-1:45pm***	***LUNCH: 12:3O-2:00pm***
Arrival of the participants **5:00-6:30 pm** **6:30 pm** ***John Bateman*** *Is the signalling of text organisation a transmodal phenomenon?*	**Thomas, M., Delin J. & Waller, R.** A framework for corpus-based analysis of the graphic signalling of discourse structure **Longrée, D. & Mellet, S.** Analysis of Textual Data, some topological methods for studying text structure indicators: the case of Latin historic narratives **Adam, C. & Vergez-Couret, M.** Signalling Elaboration: Combining Gerund Clauses with Lexical Cues ***Coffee Break: 4:00-4:20*** **Lavid, J., Arús J. & Moratón, L.** Signalling genre through Theme: The case of news reports and commentaries **Fabre, C., Ho-Dac, L.-M., Péry-Woodley, M.-P. & Rebeyrolle, J.** On the signalling of multi-level discourse structures **Oakey, D. & Mathias, P.** Lexico-grammatical Discourse Features of Interdisciplinary and Interprofessional Co-operation	**Laippala, V.** *0... Second... Finally...* Marking and unmarking of items in sequential text organisation **Giraud, S. & Thérouanne, P.** Role of lexico-syntactic and prosodic cues in spoken comprehension of enumerations in sighted and blind adults **3:15-6:30 pm** ***Guided tour to Moissac*** *Our tour guide takes you to the ancient St Peter monastery to discover the magnificent romanesque cloister, consecrated in 1100. Then, visit of the church and its impressive porch, dedicated to the Apocalypse. Through the city, the tour ends with a walk towards the Pont Canal, surprising monument where the Canal of Garonne spans the Tarn River.*	Departure of the participants
DINNER: 7:30 pm	***DINNER: 7:30 pm***	***DINNER: 7:30 pm***	

Programme : Lydia-Mai Ho-Dac, Julie Lemarié, Marie-Paule Péry-Woodley and Marianne Vergez-Couret. Proceedings: Lydia-Mai Ho-Dac. Administration: Julie Lemarié. Accommodation: Marianne Vergez-Couret. Excursion: Anne Le Draoulec.

http://w3.workshop-mad2010.univ-tlse2.fr/