Pronominalization revisited[*]

Renate Henschel and **Hua Cheng** and **Massimo Poesio**
HCRC, University of Edinburgh, UK
{henschel,huac,poesio}@cogsci.ed.ac.uk

Abstract

Pronominalization has been related to the idea of a local focus – a set of discourse entities in the speaker's centre of attention, for example in Gundel et al. (1993)'s givenness hierarchy or in centering theory. Both accounts say that the determination of the focus depends on syntactic as well as pragmatic factors, but have not been able to pin those factors down. In this paper, we uncover the major factors which determine the focus set in descriptive texts. This new focus definition has been evaluated with respect to two corpora: museum exhibit labels, and newspaper articles. It provides an operationalizable basis for pronoun production, and has been implemented as the reusable module `gnome-np`. The algorithm behind `gnome-np` is compared with the most recent pronoun generation algorithm of McCoy and Strube (1999).

1 Introduction

Besides the well established problem of pronoun resolution, pronoun generation is now attracting renewed attention. In the past, generation systems generated pronouns without attaching much importance to the problem, one notable exception being the classical algorithm of Dale (1990), loosely based on centering theory. With the emergence of corpus based studies in computational linguistics, the question arises whether it is possible to refine known standard algorithms, or whether an improvement is only to be achieved with the help of world knowledge reasoning – a matter too complex to be dealt with reliably at this time. The former direction is represented by the pioneering work of McCoy and Strube (1999). They propose a refined algorithm for the choice between definite description on the one hand and pronoun on the other for animate referents[1] , which is based on distance, time structure and ambiguity constraints.

Here we introduce a more general algorithm for the pronominalization decision that is valid not only for animate but for inanimate referents as well. In conformity with McCoy and Strube, we group noun phrases with definite determiner and proper names together under the term "definite description". The algorithm proposes a new pronominalization strategy, which beyond McCoy and Strube (1999)'s criteria makes use of the discourse status of the antecedent and parallelism effects.

The algorithm has been implemented as the reusable module `gnome-np`. It has been re-used in the web hypertext generation system ILEX (see Oberlander et al. (1998)). It shows an accuracy over 87% with respect to two corpora (each 5000 words) of different genres.

2 Accounts of pronominalization

In previous accounts pronominalization has been related to the idea of a local focus of attention: a set of discourse referents who/which is in the center of attention of the speaker (e.g. Sidner (1979), givenness hierarchy (Gundel et al., 1993), centering theory (Grosz et al., 1995), RAFT/RAPR (Suri, 1993)). Whereas (Gundel et al., 1993) do not attempt to make their focus notion operationalizable, this has been attempted by further developments of centering. However these have mostly been applied to the pronoun resolution problem. In the following we discuss three versions of centering and show that their application to the pronoun generation problem is nevertheless limited.

Centering. Centering was developed to explain local discourse coherence; the extent to which it benefits pronoun generation is however not immediately clear. In centering,

[*] The work reported in this paper has been carried out with the financial support of UK ESPRC grant L51126.

[1] We use the terms "discourse entity" and "referent" synonymously in this paper.

the discourse entities evoked in a certain utterance u_i are called forward-looking centers (Cfs). It is assumed that they are partially ordered. As a major determinant of the ordering, the grammatical function hierarchy (roughly: SUBJ>OBJ>OTHERS) has been proposed. Because other factors affecting the order have not been elaborated in detail, this ranking (as the only operationalizable handle) has become the standard ranking in several computational applications of centering. The backward-looking center (henceforth Cb) is a distinguished member of the Cfs, which is defined as the most highly ranked member of the Cfs of the previous utterance u_{i-1} which is realized in u_i. The Cb is considered as the local focus of attention. Centering states two rules. Only the first rule makes a claim about pronominalization: If any element of the utterance u_{i-1} is realized in u_i as pronoun, then the Cb must be pronominalized in u_i as well. As noted by McCoy and Strube (1999), this rule applies only in the case that two subsequent utterances share more than one referent, and that the non-Cb referent is pronominalized in the second utterance. But why this non-Cb referent is realized as a pronoun is not given by the theory.

However, following more the spirit of centering than the actual definition, one can understand the Cb as the referent which is preferably pronominalized. General pronominalization of the backward-looking center was in fact a claim of early centering, but had to be abandoned because of counter-evidence from real discourse. But the idea that pronominalization of the Cb could be a means of establishing local discourse coherence is still prevalent. It has accordingly been used by some generation systems to control pronominalization – e.g. in the ILEX system (Oberlander et al., 1998), the Cb is always realized as a pronoun.

Semantic centering. Centering is also found in Dale (1990) as the method of pronominalization control. However, Dale's center definition differs from standard centering theory in that it is defined semantically and not on the basis of a syntactic ranking.[2] This approach has some appeal, especially for generation, because it supports the natural modularity between strategic generation – which would determine the semantic center for each utterance – and tactical generation – which decides about grammatical functions.

Functional centering. Finally, the centering version suggested by Strube and Hahn (1999) appears to reveal an underlying discourse mechanism responsible for centering: the information structure of an utterance (roughly the given-new pattern) is the deeper reason for the ranking of the forward-looking centers. This permits a generalization of standard centering into a language-independent theory covering both free and fixed word-order languages. It is however then surprising that this result is not made maximal use of in the subsequent generation-oriented work of McCoy and Strube (1999).

Beyond centering. The questions which remain open with all three approaches - standard centering, semantic centering and functional centering - are:

$\boxed{\mathbf{P1'}}$ Why are in real texts a large number of Cb's not pronominalized?

$\boxed{\mathbf{P2'}}$ Why are non-Cb referents pronominalized?

or expressed independently of centering:

$\boxed{\mathbf{P1}}$ Why are in real texts a large number of discourse entities with an antecedent in the previous utterance not pronominalized?

$\boxed{\mathbf{P2}}$ Why can more than one entity be pronominalized in one utterance?

From a corpus-driven view, question $\boxed{\mathbf{P1}}$ is the larger problem.

McCoy and Strube (1999) were the first to suggest an algorithm for generation which solves these problems. It was motivated by the observation that a large percentage of NPs which would have been realized by pronouns using known algorithms, are in fact *not* realized as pronouns in real text. They suggest that such NPs serve to mark 'time changes' in the discourse. Their algorithm accordingly makes use of distance, context ambiguity and temporal discourse structure to decide about pronominalization. In our work, we have considered a corpus of a different genre in which temporal change does not play a determining role: descriptive texts. We propose a new algorithm

[2] In particular, Dale adopts the result of the action denoted by the previous clause of a recipe as the center.

that significantly simplifies the problem of pronoun choice. It is based on a new definition of the local focus, which views the discourse status of the antecedent as the major motivation behind focusing. The algorithm performs equally well when applied to McCoy and Strube's corpus of newspaper articles.

3 Corpus analysis

The algorithm we will present below has been developed in close relation to the MUSE corpus – a corpus of museum exhibit labels[3]. The corpus is a collection of web pages of the Paul Getty Museum, pages from an exhibition catalogue, and pages from a jewellery book. Typical characteristics are the central role of inanimate referents in these texts, and the lack of temporal change – thus providing an interesting counterpart to the newspaper genre investigated by McCoy and Strube.

With an overall set of around 5000 words, the corpus contains 1450 NPs. Each NP has been annotated with respect to, among others, grammatical function, discourse status, gender, number, countability, and antecedent relationships. 23% of the NPs form reference chains (i.e. at least two mentions of one and the same referent in one text), the other 77% are only mentioned once. We have 101 different reference chains; the chain-forming NPs fall into 101 discourse-new and 213 anaphoric NPs. In the following, we will only discuss the anaphoric NPs. 50% of the anaphoric NPs are realized as definite descriptions, 50% as pronouns. We distinguish between locally bound pronouns, which are determined syntactically (Binding Theory, (Chomsky, 1981)), and which we expect the tactical generator to handle correctly, and pronouns which are not locally bound – so-called discourse pronouns. We investigated possible correlations between the discourse pronouns and semantic/pragmatic features of their context.

The basic notions that we found were distance, discourse status of the antecedent, and grammatical function of the antecedent. All three notions need a precise definition.

Distance. To be able to determine the distance between a discourse entity and its antecedent, a precise determination of what counts as utterance unit is necessary. Following Kameyama (1998), we take as **utterance** unit the finite clause. Relative clauses and complement clauses are not counted as utterances on their own. This means that we count clauses containing complement clauses or relative clauses as single utterances.[4,5] The **previous utterance** is the preceding utterance at the same level of embedding.

Note that we allow the treatment of clauses with VP coordination (subject ellipsis) as complex coordinated clauses as done in Kameyama (1998), thus handling subject ellipsis as a discourse pronoun; our algorithm does not insist on this view however.

The following correlation between pronoun use and distance was found in our corpus: 97% of the pronouns have an antecedent in the same or the previous utterance.

Discourse status. The information status of a discourse entity in an utterance is either *given* or *new*. We use these terms with an identical meaning as *ground* and *focus* in Vallduvi (1993). Discourse status, as introduced by Prince (1992), is a similar but different notion: A discourse entity is **discourse-old**, if it has been mentioned before in the same discourse; it is **discourse-new** otherwise. All cases of givenness by indirect means like part-whole, set-member relationships, other bridging relations, inferences (Prince's inferrables, anchored and situationally evoked entities) are judged as discourse-new, thus taking into account only the identity antecedent relationship. We share Prince's opinion that pronominalization has to do with discourse status, whereas definiteness has to do with information status.

66% of all short-distance discourse pronouns in the MUSE corpus refer to an antecedent which is in itself discourse-old.

Subjecthood. The third strong correlation is the relation between pronoun use and the grammatical function of the antecedent. 63% of discourse pronouns have a subject as antecedent. The following table shows the overall distribution of antecedent properties for short-distance

[3] URL: http://www.hcrc.ed.ac.uk/~gnome/corpora

[4] This deviates from Kameyama, who analyzes reported speech as separate utterance.

[5] Complement and relative clauses consisting of more than one finite clause create their own internal level of focusing.

discourse pronouns and (shown in brackets) for short-distance definite descriptions.

	old	new
subject	38% (22%)	25% (12%)
not subj	28% (18%)	9% (48%)

4 Algorithm

Based on these corpus study results, we define a new notion of the **local focus** – the set of referents which are available for pronominalization. The local focus is updated at each utterance boundary, and is defined as the set of referents of the previous utterance which are:

(a) discourse-old, or

(b) realized as subject.

This set can theoretically contain more than one referent, but in most cases, (a) and (b) are one and the same singleton set, which could be seen as the well-known Cb. Thus standard centering appears as a special case of our approach. This account means that newly introduced referents are not immediately pronominalized in the following utterance, unless they have been introduced as subject – an observation made by Brennan (1998) and now confirmed with respect to our data also.

The proposed definition of the local focus generalizes the focusing mechanisms assumed in centering and introduces the discourse status of the antecedent as one main criterion behind the pronominalization decision. It is interesting to note that McCoy and Strube (1999) also make use of the discourse status of the antecedent without mentioning it explicitly. For a certain subset of intrasentential anaphoric relations in ambiguous contexts they propose pronominalization in case the antecedent would be the preferred one in Strube (1998)'s pronoun resolution algorithm. Because the set of antecedents is ranked there with respect to information status, this is identical with our proposal. Why they do not use the discourse status as a general criterion is not clear. We believe that the discourse status of the antecedent as pronominalization trigger is a general rule in discourse semantics.

The central role of discourse status and subjecthood are in our opinion not accidental. The two notions reflect two typical strategies to introduce a new referent into the discourse. We will assume here the unmarked information structure of an utterance: *given* — *new*. The subject usually is part of (or identical to) the *given*. Let X be a certain referent which is newly introduced in utterance (u1), and referred to again in the following utterance (u2). In the first strategy, X is introduced in the *new* nonsubject part of (u1). And in this pattern the second mention of X in (u2) is not pronominalized. In example (1) given in Figure 1 the local focus for utterance (u2) has one element: {*he*}; *"the main rooms"* is new in (u1) and not pronominalized in (u2). The other typical strategy is where the referent is first mentioned in a subject position. This is typical for a segment onset, or the beginning of a text. Often this referent is given by other means – for example, by reference to a picture, or to a related object. In example (2) of Figure 1, the second mention is pronominalized. Thus the subject position seems to function as creating a givenness allocation for the denoted referent. These two strategies roughly correspond with two types of thematic development identified in Daneš (1974).

Parallelism. Our definition of the local focus licenses 91% (62 of 68 pronouns) of all short-distance discourse pronouns in our corpus. Looking at the pronouns violating the proposed account, we made an interesting observation: most of them occur in contexts of strong parallelism. We call an anphoric NP np_2 **parallel** if it has an antecedent np_1 in the previous utterance, and np_1 and np_2 have the same grammatical function. For work with real text, it is useful to include cases where np_2 is a possessive or genitive NP inside a certain np_3, and np_1 and np_3 have the same grammatical function. Depending on the concrete function, we distinguish subject and object parallelism. Strong parallelism is a simultaneous subject and object parallelism in two consecutive clauses. Strong parallelism always overrides the local focus criterion, and allows for pronominalization of referents with discourse-new antecedents in nonsubject position.

The local focus definition refined by the parallelism effect is an explanation for question $\boxed{\textbf{P2}}$ and a small portion of $\boxed{\textbf{P1}}$, but most cases of problem $\boxed{\textbf{P1}}$ remain open. Two reasons for not pronominalizing a referent which is a member of the local focus need to be considered:

$\boxed{\textbf{P1.1}}$ ambiguous context,

(1) (u1) *Shortly after inheriting the building in 1752, he commissioned the architect Pierre Contant d'Ivry to renovate* ***the main rooms.***
(u2) *The engravings for* ***these rooms*** *, showing the wall lights in place, were reproduced in Diderot's Encyclopaedie, one of the principal works of the Age of Enlightenment.*

(2) (u1) *Scottish born, Canadian based jeweller,* ***Alison Bailey-Smith*** *, constructs elaborate and ceremonial jewellery from industrial wire.*
(u2) ***Her*** *materials are often gathered from sources such as abandoned television sets ...*

(3) (u1) *With attachments such as an ocular micrometer,* ***the microscope*** *incorporates the latest scientific technology of the mid-1700s.*
(u2) *The design of* ***its*** *curving gilt bronze stand was the height of the Rococo style ...*

(4) (u0) *the table probably came from* ***the Trianon de Porcelaine*** *, a small house built for the King's mistress, Madame de Montespan, on the grounds of the Palace of Versailles.*
(u1) *This table's marquetry of ivory and horn, painted blue underneath , would have followed the* ***house's*** *blue-and-white color scheme, imitating blue-and-white Chinese porcelain, a fashionable and highly prized material.*
(u2) *Blue-and-white ceramic tiles decorated the* ***house,*** *...*

Figure 1: Corpus examples

P1.2 discourse structure signalling.

Ambiguity. Along with McCoy and Strube we argue that ambiguity with respect to gender/number influences the pronominalization decision: members of the local focus which have a competing referent (referent with similar gender/number) in some span to the left of the referent to be generated should not be realized as pronouns so as to minimize the inference load for the reader. However, not to allow pronominalization in **all** ambiguous context situations does not appear to be consistent with real texts (McCoy and Strube, 1999). In the MUSE corpus one third of all focal NPs occur in ambiguous contexts, one half of them is pronominalized, the other half is not. Two questions require a precise answer to use the ambiguity constraint in a generation algorithm:

- Which set of previously mentioned referents or text span is taken into account for referents to be in competition?
- Which referents are pronominalized despite an ambiguous context?

The answer is surprisingly simple: Referents of the previous utterance which are not in the local focus do *not* disturb pronominalization, even if they have the same gender/number. Only if the actual referent has a competitor in the local focus, is pronominalization blocked. This is illustrated in Figure 1 with examples (3) and (4), respectively. In (3) *the microscope* is discourse-old and the only member in the local focus for (u2); the competing referents *ocular micrometer* and *technology* are new and hence not focal for utterance (u2). In (4), the local focus for (u2) is $\{$*the table, the marquetry, the house*$\}$.

A slight improvement of the performance of the algorithm can be achieved by regarding the role of "heavy" nonrestrictive modification. Including the referents of discourse-new NPs which are amplified by appositions or nonrestrictive relative clauses into the set of possible competitors improves accuracy slightly.

Discourse structure signalling. It is now known that definite descriptions (or more general overspecified NPs) signal the start of a new discourse segment (Passonneau, 1996; Vonk et al., 1992). For most generation systems generate from an RST-like text plan, discourse segments are naturally given. The only question from the generation perspective is the degree of detail provided by the segmentation.

Our algorithm `gnome-np` assumes that the discourse segmentation has already been specified. At each segment boundary, the local focus is set to `nil`, thereby disallowing pronominalization for all discourse entities of the first utterance in the segment onset.

It is also well known that planned discourse with repeated phrases at the beginning of a clause are seen as 'bad style'. Identical repeated pronouns at the clause onset are rarely found in expository and descriptive texts (2.6% of all discourse pronouns in our corpus). Human writers usually avoid possibly dull lack of variation by employing various aggregation techniques.

```
Let X be a referent to be generated in utterance (u2), and $focus$ be the set of referents of the previous
utterance (u1) which are
    (a) discourse-old, or
    (b) realized as subject.
(1)  X has an antecedent beyond a segment boundary                          def description
(2)  X has an antecedent two or more utterances distant                     def description
(3)  X has an antecedent in (u1), and
     (3a) X occurs in strong parallel context                               pronoun
     (3b) X $\notin focus$                                                  def description
     (3c) X $\in focus$ and
          • X has a competing referent Y $\in focus$                        def description
          • X has a competing referent Y in (u1) amplified with appo-       def description
            sition or nonrestrictive relative clause
          • else                                                            pronoun
The repetition blocking rule overrides the pronominalization suggested in (3c) to a definite description.
```

Figure 2: The algorithm

Thus pronoun repetition blocking seems to be an aggregation trigger rather than a motivation for definite description generation. We hypothesize that the apparent frequency of definite descriptions in planned discourse has much to do with repetition blocking, but is used with respect to a very fine-grained, probably genre-specific discourse structure. One candidate for this is the temporal structure in newspaper articles proposed by McCoy and Strube.

When evaluating our algorithm, we only used the paragraph segmentation given in the corpus. But for generation systems, which usually are not equipped with developed aggregation modules, we have also made available a pronoun repetition blocking rule: If a discourse entity in the local focus has a nonpossessive pronominal antecedent, pronominalization will be blocked at this time. Figure 2 summarizes the algorithm.

The presented pronominalization algorithm has been implemented in the reusable module `gnome-np`. `gnome-np` consists of a component for discourse model management and one for NP form determination. It is designed to be plugged in after text planning, conceptualization, and sentence planning, but before tactical generation.

5 Evaluation

A comparison of the performance of our algorithm with the annotated MUSE corpus and McCoy and Strube's newspaper corpus is given in Table 1. The evaluation has been carried out for the algorithm `gnome-np` without employing the repetition blocking rule and without a fine-grained discourse segmentation. Layout segments were used for the MUSE corpus. Because the number of annotated segment onsets for the newspaper corpus is not easy to re-establish, we give here two figures for this corpus: first without any segment onset signalling (lower bound), and second with the assumption that 15 short-distance definite descriptions mark segment onsets. The figures include locally-bound pronouns to yield better comparability with McCoy and Strube. The figures in the columns 'gnome-np' represent those NPs whose form is predicted correctly by the new algorithm when evaluated against the annotated corpora.

The figures in Table 1 show that our algorithm performs very well in both domains, even without using a finer discourse segmentation such as temporal structure. Moreover, it performs better on McCoy and Strube's corpus than their own algorithm, which successfully predicted the choice between realization by pronoun and realization by definite description in 84.7% of all cases. The disagreements occur first for long distance pronouns (in our terminology: pronouns more than one clause distant) and, second, in longer referent chains with well established focus. For the latter, whereas `gnome-np` would always suggest a pronoun, the real discourse swaps between pronoun and definite description. Thus a finer segmentation or a repetition blocking rule could still improve the result further.

	MUSE	gnome-np	agreement	newspaper	gnome-np		agreement	
pronouns	112	101	90.2%	302	267		88.4%	
def descriptions	101	86	85.1%	225	187	202	83.1%	89.7%
total	213	187	87.8%	527	454	469	86.1%	89.0%

Table 1: Performance comparison

6 Conclusions

This paper has presented a new algorithm for the pronominalization of third person discourse entities. The algorithm, first, is implemented as a reusable module for generation systems and, second, provides a theoretical account of pronominalization in general.

The proposed algorithm provides a solution for question **P2** above by widening the definition of local focus to be a set with possibly more than one referent. The algorithm also offers a new solution for problem **P1.1** above, ambiguous pronoun generation. Discourse structuring (**P1.2**) is assumed as given. A sufficiently fine-grained discourse structuring has been explored, for example, by McCoy and Strube for their domain of newspaper articles, but remains an issue for future research for other domains. We have shown that next to proximity, the discourse status of the antecedent is a main criterion for triggering pronominalization.

The suggested algorithm generalizes known focusing accounts. Gundel et al. (1993)'s cognitive status of being "in focus" is now approximated by the set of all discourse-old entities and the subject of the previous utterance. The new focus determination is also a generalization of centering's Cb. The focus so defined serves two functions simultaneously: to trigger pronominalization, and to provide the set of competitors for pronoun generation in ambiguous contexts. Although our training corpus is too small to justify general claims, the evaluation with respect to the newspaper genre provides evidence that this finding is valid for planned discourse in general, independent of the concrete genre.

References

Susan Brennan. 1998. Centering as a psychological resource for achieving joint reference in spontaneous discourse. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, pages 227 – 250. Clarendon Press, Oxford.

Noam Chomsky. 1981. *Lectures on government and binding.* Foris, Dordrecht.

Robert Dale. 1990. *Generating referring expressions.* The MIT Press, Cambridge, Massachusetts.

František Daneš. 1974. Functional sentence perspective and the organisation of the text. In František Daneš, editor, *Papers on Functional Sentence Perspective*, pages 106–128. Academia, Prague.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203 – 164.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69:274 – 307.

Megumi Kameyama. 1998. Intrasentential centering: A case study. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering Theory in Discourse*, pages 89 – 114. Clarendon Press, Oxford.

Kathleen McCoy and Michael Strube. 1999. Generating anaphoric expressions: Pronoun or definite description? In *Proceedings of ACL '99 Workshop: Reference and discourse structure*, pages 63 – 71.

J. Oberlander, M. O'Donnell, A. Knott, and C. Mellish. 1998. Conversation in the museum: experiments in dynamic hypermedia with the intelligent labelling explorer. *New Review of Multimedia and Hypermedia*, pages 11 – 32.

Rebecca Passonneau. 1996. Using centering to relax gricean constraints on discourse anaphoric noun phrases. *Language and Speech*, 39(2):229 – 264.

Ellen F. Prince. 1992. The ZPG letter: Subjects, definiteness and information status. In W. C. Mann and S. A. Thompson, editors, *Discourse desciption: Diverse linguistic analyses of a fund-raising text.* John Benjamins, Amsterdam.

Candace L. Sidner. 1979. *Towards a computationally theory of definite anaphora comprehension in English disourse.* PhD thesis.

Michael Strube and Udo Hahn. 1999. Functional centering – grounding referential coherence in information structure. *Computational Linguistics*, 25(3):309 – 344.

Michael Strube. 1998. Never look back: An alternative to centering. In *Proceedings of Coling-ACL '98*, pages 1251 – 1257.

Linda Z. Suri. 1993. *Extending focussing frameworks to process complex sentences and to correct the written English of proficient signers of American Sign Language.* PhD thesis.

Enrico Vallduvi. 1993. Information packaging – a survey. Technical report, HCRC research paper RP-44.

W. Vonk, G. Hustinx, and W. Simons. 1992. The use of referential expressions in structuring discourse. *Language and Cognitive Processes*, 7(3/4):301 – 333.