A Characterization of Underspecified Anaphora and its Consequences on the Annotation of Anaphoric Relations

When resolving anaphora, the precise identification of the antecedent is sometimes difficult. In a first set of cases linked to abstract anaphora (*"in **that context**"*, *"**it** happens"*), the reason lies in the exact delimitation of the antecedent (previous verbal phrase? previous sentence? whole paragraph?). In a second set of cases, several concrete antecedents are possible and the hearer or reader cannot decide between them. Moreover, deciding is not mandatory to the comprehension of the utterance or discourse, and the antecedent can stay underspecified. This phenomenon can be called 'underspecified anaphora', or anaphora with 'fuzzy' or 'vague' antecedent. We illustrate it with French data, and we propose a first classification of types of potential underspecified antecedent. We then deduce principles for their annotation.

A characterisation of underspecified anaphora in French We focus on the classical French personal pronouns *"**il**"* and *"**elle**"* (Kleiber, 1994) when they are ambiguous. To simulate the ambiguity in the English version, two different pronouns are sometimes used.

(1) « Je lui ai dit sur un ton de plaisanterie que son idée était intéressante, qu'**elle** montrait les choses sous un angle auquel en effet on n'est pas habitué […] » Frantext corpus, K899.
"I told her that her idea was interesting, that **she/it** shows things in an unexpected way."

(2) « […] dans le comportement de tout public quel qu'**il** soit » Frantext corpus, R875.
"the behaviour of the public, whatever **he/it** is."

(3) « […] les rapports des rapporteurs et les documents auxquels **ils** se réfèrent, pour autant qu'**ils** n'ont pas été communiqués antérieurement, sont […] » Frantext corpus, P659.
"the reports from the reviewers and the related documents, if **they** have not been […]."

(4) « […] cette faille, cette malédiction qui est comme une résurgence laïque du péché originel, est d'autant plus grave qu'**elle** laisse un vide […] ».
"this flaw, this curse is all the more serious since **it** leaves a void."

(5) « Voulez-vous Jean Dupont comme époux ? – Oui, je **le** veux ».
"Do you want to take John Dupont to be your husband? – Yes, I want **him/it**."

(6) « Jean a coupé tout le bois. Il **l**'a fait en petits morceaux ».
"John cut wood. He did/made **it** in small parts."

These examples put forward nominal phrases like *"le N_1 de le N_2"* as well as possessives, coordinations and juxtapositions. Attributing a precise antecedent is each time not really crucial for the comprehension of the text. In (1), the person as well as her idea can show things in an unexpected way. Even if the two semantic representations differ, they convey the same global signification, since the person can be fused to her idea in the context of this sentence. In (2) and (3), a deep analysis can be made to identify the referent the writer or speaker had in mind, but the result is not guaranteed and the underspecified solution may be better. In (4), two possible interpretations are the precision and the correction. The juxtaposition gives no indication to choose, and both stay relevant. In some other unambiguous examples, the antecedent is the first noun: "*le rêve, l'évasion est nécessaire. **Il** permet de mieux supporter les tracas de la vie quotidienne*" (*"dreaming, daydreaming is necessary. It allows feeling better"*). In (5) and (6), the ambiguity implies a concrete antecedent and an abstract one. Following the point of view of (Prandi, 1987) concerning some syntactic ambiguities, we can say that underspecified anaphora is more frequent than we think (in French in any case), since there is often no simple linguistic way to avoid the ambiguity. With a natural language processing point of view, we can notice that classical algorithms for anaphora resolution (Mitkov et al., 2007), and for instance salience-based methods (Miltsakaki, 2007) are useless face to this phenomenon.

Confronting these examples to others from scientific studies or Frantext corpus, we propose the following characterization of potential underspecified antecedents:

1. **Possessives**, when the possessor is animated and the possession is a personality trait to which it can be assimilated, or when the possessor is an object and the possession is its main function. The key is that both of them have to be linked (one may be a facet of the other, or may belong to its set of properties).
2. **Complex nominal phrases "*le N_1 de le N_2*"**, with the same participants than in the previous situation.

3. **Coordinations**, in the same case and in cases where the sentence elements are in plural: *"the N_1 and the N_2 […] they […]"*.
4. **Juxtapositions**, in particular when it is hard to distinguish a simple enumeration from a reformulation, and, in the latter case, to distinguish a precision from a correction (example 4). Another example implies several persons and involves a possible precision: in *"the First Secretary, Mister Smith, and his wife"*, how many persons are mentioned? two or three?
5. **Evolutive referents:** in examples such as *"as a child, Marcello […], but adult Marcello is […]* ***he*** *[…]"*, does *"**he**"* refer to *"Marcello"* or to *"adult Marcello"* (or to "child Marcello" if we consider not only text spans but also semantic referents)? It is sometimes impossible and useless to make a choice.

To these cases we have to add the situations where the antecedent is not linguistically mentioned. This is the case with collective pronouns: *"**ils** ont encore augmenté les impôts"*, *"**they** have increased taxes again"* (Kleiber, 1994). This is also the case with antecedentless anaphora (Cornish, 1996).

Consequences on the annotation of anaphoric relations A way to apprehend the identification of antecedents consists of annotating (as a means to represent semantic concerns and not to evaluate computational systems). Here is a first list of annotation principles linked to underspecified anaphora:

1. The determination of an underspecified antecedent is necessary if we don't want the interpretation to be reduced to a particular or irrelevant choice.
2. **Alternatives principle**: in most of cases, several alternatives confront themselves and the choice does not matter (the person or her idea, the whole documents or only the reports, etc.). The aim is to determine what are the alternatives, and to group them as the antecedent. The initial intended antecedent from the writer or speaker is in the set of alternatives, but all of them are relevant for the reader or hearer (and for the semantic representation determination). The underspecification cannot be the source of a mistake or a misunderstanding. This principle has to be distinguished from the principle consisting of grouping some antecedents into a composite one, as in *"John was sleeping. Mary was reading.* ***They*** *were happy"*.
3. **Feasibles principle**: in some other cases, alternatives may be hard to determine, for instance with juxtapositions. The aim here is to identify all the possibilities and to label them as feasible. The intended antecedent from the speaker is in the set of feasibles, but only some of them are relevant for the semantic representation determination, i.e., underspecification may be a cause of mistake. This is the case with the example *"the First Secretary, Mister Smith, and his wife"*.
4. **Double mark-up principle**: for abstract anaphora such as *"**it** happens"*, alternatives and feasibles are not sufficient. The best solution would be to manage fuzzy or progressive limits for the antecedent, which is not compatible with mark-up constraints. We propose to manage two sets of tags, a first one corresponding to the limit inferior and the second one to the limit superior. To these tags correspond two discourse entities, and every entity in-between is acceptable. This principle differs from the 'MIN' attribute from MUC-7 (Message Understanding Conferences). The 'MIN' attribute has been designed to reduce a referential expression to a minimal mention like a proper noun, and emphasizes in this way a characteristic of a particular discourse entity. In the case of underspecified anaphora, there are several discourse entities in presence. Each of them can be characterised by a minimal mention and therefore with a 'MIN' attribute. The double mark-up principle is then independent from the 'MIN' attribute.

References

Cohen, A. (2006) Anaphora resolution and minimal models. *Proceedings of the 5th International Conference on Inference in Computational Semantics (ICOS-5)*, Buxton, England, pp. 7-16.

Cornish, F. (1996) 'Antecedentless' anaphors: deixis, anaphora, or what? Some evidence from English and French, *Journal of Linguistics*, 32, pp. 19-41.

Kleiber, G. (1994) *Anaphores et pronoms*. Duculot, Louvain-la-Neuve.

Miltsakaki, E. (2007) A rethink of the relationship between salience and anaphora resolution. *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium*, Lagos, Portugal, pp. 91-96.

Mitkov, R. *et al*. (2007) Anaphora resolution: to what extent does it help NLP applications? In: *Anaphora: Analysis, Algorithms and Applications*, Springer-Verlag, Berlin Heidelberg.

Prandi, M. (1987) *Sémantique du contresens.* Minuit, Paris.