

Italian texts annotated with co-reference and with discourse topic shifts; ii) to investigate whether there is a systematic relation between various types of referring expressions and their use in different transition states in the two languages; iii) to individuate similarities and differences in co-referential chains and in the use of referring expressions in discourse topic shifts in Danish and Italian; iv) to study whether different referring strategies are used in fiction and non-fiction texts.

The paper is organised as follows. In section 2 we discuss related work and research which have inspired us. In section 3 we shortly present our data and in section 4 we describe our annotation scheme and discuss inter-annotator agreement results. In section 5 we discuss the results of our analysis of the annotated data and finally, in section 6 we make some concluding remarks and present work still to be done.

2 Related Work

The relation between reference and discourse structure has been pointed out in numerous studies, among many (Kuno, 1972; Halliday and Hasan, 1976; Hobbs, 1979; Grosz and Sidner, 1986; Cristea and Ide, 1998). Centering is about local coherence, but the theory presupposes global coherence as proposed by Grosz and Sidner (1986). In Centering the relation between continuations or shifts in co-reference chains and the use of referring expressions is modelled in terms of so-called transition states and of the preferences holding among them. These preferences reflect the assumption that the mental effort to process reference is less when the central topic of discourse is maintained and when this topic is realised by the most salient entity than when the central topic changes or is realised by a less prominent entity. Because pronouns signal reference to the most salient entities, pronominal chains are assumed to be more frequent in center continuations than in center shifts. The transition types proposed by Brennan, Friedman, and Pollard (1987) are in (table 1). The use of other types of re-

	$C_b(U_n) = C_b(U_{n-1})$ OR no $C_b(U_{n-1})$	$C_b(U_n) \neq C_b(U_{n-1})$
$\frac{C_b(U_n)}{C_p(U_n)} =$	CONTINUE	SOFT-SHIFT
$\frac{C_b(U_n)}{C_p(U_n)} \neq$	RETAIN	ROUGH-SHIFT

Table 1: Transition states

ferring expression after the various transition states is not explored in the Centering theory. However, many researchers in the Centering framework have looked at aspects which are central to the relation between reference and discourse structure including the definition of transition states, the presence and/or uniqueness of backward-looking centers and the realisation of centers, see especially (Brennan, Friedman, and Pollard, 1987; Strube and Hahn, 1999; Fais, 2004; Poesio et al., 2004; Kibble and Power, 2004).

Independently from the Centering framework, Givón (1983) provides an analysis of the relation between topic shifts and use of subject referring expressions in English and Pidgin English monologues. He recognises two kinds of junctures in his data: minor junctures after clauses and major junctures after sentences.

In this paper we look at both global and local coherence and are especially interested in the types of nominal referring expression used in Danish and Italian texts after various transitions. We are strongly inspired by the work of Di Eugenio (1996) who analysed occurrences of Italian pronouns and full nominal phrases in subject position respect to a particular version of the Centering’s transition states. Her focus was on the use of zero pronouns².

We adopt a combination of two cognitive models of referring expressions: the GIVENNESS HIERARCHY proposed by Gundel, Hedberg, and Zacharski (1993) and Ariel (1988), (1994)’s ACCESSIBILITY MARKER SCALE.

Gundel, Hedberg, and Zacharski (1993) organise the assumed cognitive statuses of discourse entities in their GIVENNESS HIERARCHY and connect each status to a precisely identified referring expression, exemplified by an English nominal phrase (table 2). They argue for the universality of their hierarchy, although they notice that not all languages have referring expressions for each status in the hierarchy. The GIVENNESS HIERARCHY is interesting because, differing from related cognitive models, it assumes that the various cognitive statuses are implicationaly related and not mutually exclusive. Thus, according to this theory, a referring form encodes the necessary and sufficient status it belongs to as well as all the higher statuses in the hierar-

²Italian is a subject pro-drop language.

in focus	>	activated	>	familiar	>	uniquely identifiable	>	referential	>	type identifiable
<i>it</i>		<i>that</i> <i>this</i> <i>this N</i>		<i>that N</i>		<i>the N</i>		indefinite <i>this N</i>		<i>a N</i>

Table 2: The Givenness Hierarchy

chy (the statuses on its left). This accounts for cases in discourse where a speaker uses a referring expression signalling a less given cognitive status than required by the context, e.g. to emphasise some entities.

One problem with the GIVENNESS HIERARCHY is that it does not account for differences between types of referring expression which do not occur in English. This is the case for the Italian zero anaphora and clitics.

A more fine-grained hierarchy of nominal referring expressions is presented by Ariel (1994). Also Ariel points out that speakers code how accessible a referent is to the addressee by using different referring expressions. Analysing the distance between antecedent and referring expressions, one of the factors that determine the accessibility of these expressions, Ariel builds up an *accessibility marker* system for referring expressions. In her system *unmarked* means prototypical, while the concept of *markedness* presupposes the notion of *formal complexity* and is connected with structural complexity, low frequency and cognitive complexity. A simplified version of Ariel’s ACCESSIBILITY MARKING SCALE (Ariel, 1994) is given in figure 1. The accessibility of the expressions de-

zero < reflexives < cliticised pronouns < unstressed pronouns < stressed pronouns < stressed pronouns + gesture < proximal demonstrative (+ NP) < distal demonstrative (+ NP) < proximal demonstrative + NP + modifier < distal demonstrative + NP + modifier < first name or last name < definite description < full name
--

Figure 1: Ariel’s Accessibility Marking Scale

creases from left to right: the highest accessibility markers being the most unmarked linguistic expressions. Thus the symbol < in the scale refers to the degree of markedness. The more (lexically) informative, the more rigidly

(unambiguously) and/or the less attenuated the form (longer or louder) of a referring expression the lower accessibility it marks.

We use Ariel’s classification of referring expressions, but assume with Gundel, Hedberg, and Zacharski (1993) that the cognitive statuses related to the different referring expressions are implicationaly related.

3 The data

We have annotated the following Danish and Italian data:

- Parallel texts: i) European law texts (7,631 running words in Italian and 7,101 running words in Danish); ii) Italian stories by Pirandello (9,018 words) and their Danish translations (9,933 words)
- Comparable texts: i) Financial newspapers: the Italian *Il Sole 24 Ore* (6,964 words) and the Danish *Børsen* (3,325 words)

The source language of the European texts is not known, but it is probably English or French.

The parallel texts and some of the comparable texts which we have annotated belong to the MULINCO corpus (Maegaard et al., 2006). Part of these texts are freely available.

In order to obviate some of the problems connected with the use of translated texts³ we have annotated articles from financial newspapers in the two languages describing similar events and written in the same period of time. Although these articles are covered by copyright restrictions, they can be obtained by the publishing editors for research.

³One of these problems is the use of referring expressions in the target language being influenced by the referring expressions used in the source language. Examples of these influence are in (Navarretta, 2007).

4 The annotation

Co-referential and referential chains in the corpus have been annotated using an extension of the MATE/GNOME annotation scheme (Poesio, 2004). Bridging anaphora have not been annotated. We use the markables proposed in the MATE/GNOME scheme, i.e. DE to mark discourse entities and SEG to annotate non nominal referring expressions. The markable LINK marks the relation between referring expressions and their antecedents.

We have added a number of attributes to these markables to encode the following information: a) the type of referring expression comprising the pronominal and nominal types recognised by Ariel (1994); b) the syntactic type of the antecedent including nominal and non-nominal antecedents, such as predicates in copula constructions, verbal phrases, clauses and discourse segments; c) the pronominal function, such as cataphoric, individual anaphoric, deictic, pleonastic, abstract anaphoric.

Only two types of relation between referring expressions and antecedents are used: *identity* and *non-identity*. The *identity* relation is used for co-reference, while *non-identity* is used for all other cases, comprising the relations between antecedents and anaphora referring to different semantic types of entity, and the relation connecting appositions to the nominal phrases they define or modify. Example 1 contains the annotation of the two appositions in the text segment *Lina Sarulli, prima Lina Taddei, ora Lina Fiorenzo* (Lina Sarulli, previously Lina Taddei, now Lina Fiorenzo) from Pirandello's story *La buon' anima*. The two appositions are bound to the proper *Lina Sarulli* by a non-identity relation.

We have added some markables to the MATE/GNOME scheme to mark pleonastic pronouns and pronouns in abandoned utterances⁴. Possessive pronouns and deictic pronouns in direct speech are also annotated. These occurrences of deictic pronouns are in most cases part of the co-referential chains in the fiction data.

Two slightly different annotation schemes are used for Danish and Italian, accounting for language specific differences, such as the fact that Italian is a subject PRO-

⁴These occur in direct speech in our fiction data.

drop language and has both independent and clitic pronouns. A kind of SEG markers, SEG1 is used to mark verbal phrases containing one or more clitic pronouns, as illustrated in example 2 where the verb form *promettendoglielo* (promising it to him) contains two clitic pronouns *gli* (to him) and *lo* (it), which co-refer with two entities whose identifiers are *n150* and *i24* respectively (*promettendo[gli]_{n150}e[lo]_{i24}*).

The data we have annotated with coreference had been previously annotated with abstract pronominal anaphora information in the DAD project. These anaphora are third-person singular pronouns whose linguistic antecedents are predicates in copula constructions, verbal phrases, clauses and discourse segments. The annotation specific to abstract anaphora is described in (Navarretta and Olsen, 2008) and comprises the semantic type of abstract referents, partially inspired by the classification of abstract objects by Asher (1993).

Discourse topics have been annotated using a variation of the annotation proposed by Rocha (2000) who distinguishes among discourse topics, segment topics and subsegment topics in English and Portuguese dialogues.

In our data paragraphs correspond in most cases to discourse segments, see (Grosz and Sidner, 1986). Discourse segments have been further divided into subtopics and subsubtopics.

A subset of the data has been marked with the transition types proposed in (Brennan, Friedman, and Pollard, 1987)⁵. The salience model adopted for annotating transition states in both Danish and Italian is mainly that proposed in (Navarretta, 2002; Navarretta, 2005) (figure 2).

We have used PALinkA (Orăsan, 2003) as annotation tool.

The first 4000 words of the Italian data were annotated by four annotators and inter-annotator agreement was automatically calculated on these data in terms of weighed *kappa* statistics⁶ (J.Cohen, 1968) using PRAM⁷. The obtained results varied from 0.60 to 0.95, depending on the type of

⁵Only the author annotated this information.

⁶Other evaluation methods are discussed by Arstein and Poesio (2008).

⁷<http://www.geocities.com/skymegsoftware/pram.html>.

- (1) <de ID="n643" firstm="MNO" syn-type="PR">
 <link Ltype="ident" POINT-BACK="n334"/>
 <W id="w2.24.15" lemma="lina" pos="NPR">Lina</W>
 <W id="w2.24.16" lemma="sarulli" pos="NPR">Sarulli</W></de>
 <W id="w2.24.17" lemma="," pos="PON">,</W>
 <W id="w2.24.18" lemma="prima" pos="ADV">prima</W>
 <de ID="n644" firstm="MNO" syn-type="PR">
 <link Ltype="no_ident" POINT-BACK="n643"/>
 <W id="w2.24.19" lemma="lina" pos="NPR">Lina</W>
 <W id="w2.24.20" lemma="taddei" pos="NPR">Taddei</W></de>
 <W id="w2.24.21" lemma="," pos="PON">,</W>
 <W id="w2.24.22" lemma="ora" pos="ADV">ora</W>
 <de ID="n645" firstm="MNO" syn-type="PR">
 <link Ltype="no_ident" POINT-BACK="n643"/>
 <W id="w2.24.23" lemma="lina" pos="NPR">Lina</W>
 <W id="w2.24.24" lemma="fiorenzo" pos="NPR">Fiorenzo</W></de>
- (2) <seg1 ATYPE="indiv" ID="i25" PTYPE="lo-clitico" syn-type="V">
 <link Ltype="ident" POINT-BACK="i24"/>
 <seg1 ATYPE="indiv" ID="i151" PTYPE="gli-clitico" syn-type="V">
 <link Ltype="ident" POINT-BACK="n150"/>
 <W id="w25.57.60" lemma="promettere" pos="VER:geru">promettendoglielo</W></seg1></seg1>

markable. The worse results were obtained in the annotation of discourse segment antecedents of abstract substantives. Examples of these abstract referring expressions are *tali situazioni* (such situations) and *questa discussione* (this discussion). Inter-coder agreement for the annotation of pronominal abstract anaphora was not calculated because it had been tested in the DAD project (Navarretta and Olsen, 2008).

An annotation example is in 3. The annotated text segment is [*La Acqua Marcia*]_i può evitare il fallimento. [*La finanziaria di Vincenzo Romagnoli*]_j ... ([*La Acqua Marcia*]_i can avoid bankruptcy. [[*Vincenzo Romagnoli*]_j's investment company]_i) [*Il Sole 24 ore*](31.12.1992)].

The annotation of co-reference is expressed by saying that the nominal phrase *Vincenzo Romagnoli's investment company*, is related to the proper *La Acqua Marcia* by an identity relation.

5 Results

The number of markables annotated in the data are given in table 3. To these markables must be added the SEG elements which code the non-nominal antecedents of abstract anaphora, pleonastic and abandoned occurrences of pronouns. The length of co-referential chains varies consistently from text type to text type independently from the analysed language. The (co)referential chains

	Zero	Clit	PRO	Name	NPs
it	1225	240	1075	762	1995
da	-	-	2331	602	1524

Table 3: Number of markables

in literary texts are much longer than those in non-literary texts. This is not surprising because the stories are longer than the financial articles and they focus on fewer subjects (persons, objects) than the analysed European texts.

In our data there are nearly 5 times more pronouns pr. 1000 words in literary data than in non-literary texts. Reference by substantives was on the contrary higher in the non-literary texts than in the literary data (here the proportion pr. 1000 words was 4 to 1).

The average distance in terms of sentences between referring expressions and their antecedents is higher in literary data than in non literary data. We have not investigated yet whether there is a relation between referential distance and number of discourse entities and possible candidate antecedents in the involved texts.

Inferable entities are more often anchored to known entities by genitives in Danish than in Italian. An example is in 4.

- (4) *Fin dal primo giorno, Bartolino Fiorenzo s'era sentito dire dalla promessa sposa...* (the fiancée)
Fra første dag havde Bartolino

Fiorenzo *hørt sin tilkommende sige...* (his fiancée)
 (From the very first day Bartolino Fiorenzo had heard his fiancée say...) Pirandello: *La buon' anima*

In Italian the distal demonstrative determiners *quel/quello/quella* (that) and *quelli/quelle* (those) followed by a substantive are used if i) there are other clauses or nominal phrases in-between the referring expression and antecedent; ii) there is temporal or spatial distance from the antecedent. In Danish the proximal demonstrative determiners *denne/dette/disse* (this/these) are used in the same contexts: *quella donna* (that woman)/*denne kvinde* (this woman); *quella sciagura* (that calamity)/*denne ulykke* (this calamity). Only if the antecedent is the immediately preceding discourse segment the proximal demonstrative determiners are used in both languages.

As noticed in (Navarretta, 2007; Navarretta and Olsen, 2008) abstract substantives are used in Italian in most cases where Danish uses abstract pronouns.

The analysis of the relation between transition states and types of referring expressions in the three stories by Pirandello is given for Italian in figure 3 and for Danish in figure 4. The figures give a scale of the significantly most frequent referring expressions occurring as centers after the various Centering transition types⁸. The results in the figures only partially confirm existing classifications of the givenness or salience of referring expressions and reflect some of the differences between Danish and Italian that we have previously discussed. An interesting fact, which cannot be seen in the figures is that in these particular data deictic pronouns are in 96% "locally" deictic and have thus been linked to the local co-referential chains. Because the amount of our data is not large, the present results are only preliminary.

6 Conclusion

We have presented a rich annotation of (co)-referential chains in Danish and Italian comparable and/or parallel data and we have dis-

⁸In the two figures *Def N. anchored* refer to all definite nominal phrases which are bound to entities previously introduced in discourse (Prince, 1981) via e.g. genitive phrases, propositional phrases, relative clauses.

cussed some dissimilarities in the use of referring expressions in the two languages. The relation between types of referring expression used to refer to the backward-looking center after different types of transition have been studied in the fiction data. Although the results are interesting they can only be considered preliminary because of the limited amount of data. Furthermore the analysed Danish texts are translations of the Italian stories, thus more differences in-between the two languages might be found in comparable data. However we believe that the strategy of looking at the relation between transition types and types of referring expression is very useful especially if conducted on more languages and on more types of text.

Currently we are annotating the transition types on the remaining data and we plan to extend our analysis to the referential distance and to the number of competing antecedent candidates.

References

- Ariel, M. 1988. Referring and accessibility. *Journal of Linguistics*, 24(1):65–87.
- Ariel, M. 1994. Interpreting anaphoric expressions: a cognitive versus a pragmatic approach. *Journal of Linguistics*, 30(1):3–40.
- Arstein, R. and M. Poesio. 2008. Inter-coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.
- Asher, N. 1993. *Reference to Abstract Objects in Discourse*, volume 50 of *Studies in Linguistics and Philosophy*. Kluwer Academic Publishers, Dordrecht, the Netherlands.
- Brennan, S. F., M. W. Friedman, and C. J. Pollard. 1987. A Centering Approach to Pronouns. In *Proceedings of ACL 87*, pages 155–162, California, USA. Stanford University.
- Cristea, D. and N. Ide. 1998. Veins theory: A model of global discourse cohesion and coherence. In *Proceedings of COLING/ACL 98*, pages 281–285, Montreal.
- Di Eugenio, B. 1996. The discourse functions of Italian subjects: a centering approach. In *Proceedings of COLING 96*, pages 352–357, Copenhagen, Denmark. Centre for Language Technology.

- Fais, L. 2004. Inferable centers, centering transitions and the notion of coherence. *Computational Linguistics*, 30(2):119–150.
- Givón, T., editor. 1983. *Topic Continuity in Discourse: A Quantitative Cross-Language Study*. John Benjamin, Amsterdam.
- Grosz, B., A. K. Joshi, and S. Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–225.
- Grosz, B. J. and C. L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–284.
- Gundel, J. K., N. Hedberg, and R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.
- Halliday, M. and R. Hasan. 1976. *Cohesion in English*. Longman, London.
- Hobbs, J. R. 1979. Coherence and Coreference. *Cognitive Science*, 3(1):67–90.
- J. Cohen. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.
- Kibble, R. and R. Power. 2004. Optimizing Referential Coherence in Text Generation. *Computational Linguistics*, 30(4):401–416.
- Kuno, S. 1972. Functional sentence perspective. *Linguistic Inquiry*, 3:269–320.
- Lambrecht, K. 1994. *Information structure and sentence form - Topic, focus and the mental representations of discourse referents*, volume 71 of *Cambridge Studies in Linguistics*. Cambridge University Press.
- Maegaard, B., L. Offersgaard, L. Henriksen, H. Jansen, X. Lepetit, C. Navarretta, and C. Povlsen. 2006. The MULINCO corpus and corpus platform. In *Proceedings of LREC-06*, pages 2148–2153, Genova.
- Navarretta, C. 2002. *The use and resolution of Intersentential Pronominal Anaphora in Danish Discourse*. Ph.D. thesis, University of Copenhagen, February.
- Navarretta, C. 2005. Combining information structure and centering-based models of salience for resolving danish intersentential pronominal anaphora. In A. Branco, T. McEnery, and R. Mitkov, editors, *Anaphora Processing. Linguistic, cognitive and computational modeling*, volume 263 of *Current Issues in Linguistic Theory*. John Benjamins Publishing Company, pages 329–350.
- Navarretta, C. 2007. A contrastive analysis of abstract anaphora in danish, english and italian. In A. Branco, T. McEnery, R. Mitkov, and F. Silva, editors, *Proceedings of DAARC 2007*, pages 103–109. Centro de Linguistica da Universidade do Porto, March.
- Navarretta, C. and S. Olsen. 2008. Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of LREC-2008*, Marrakesh, Morocco, May.
- Orăsan, Constantin. 2003. PALinkA: a highly customizable tool for discourse annotation. In *Proceedings of the 4th SIGdial Workshop*, pages 39 – 43, Sapporo, Japan, July, 5 -6.
- Poesio, M., R. Stevenson, B. Di Eugenio, and J. Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Computational Linguistics*, 30(3):309–364.
- Poesio, Massimo. 2004. The mate/gnome proposals for anaphoric annotation, revisited. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop*, pages 154–162, Cambridge, Massachusetts, USA, April 30 - May 1. Association for Computational Linguistics.
- Prince, E. F. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*. Academic Press, pages 223–255.
- Rocha, M.A.E. 2000. A corpus-based study of anaphora in english and portuguese. In S.P Botley and T. McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora*. Benjamins Publishing Company, pages 81–94.
- Strube, M. and U. Hahn. 1999. Functional Centering - Grounding Referential Coherence in Information Structure. *Computational Linguistics*, 25(3):309–344.

**FOCUS PROPER < SUBJECT < OBJECT/PrepOBJECT < OBJECT2 < OTHER
COMPLEMENTS < ADJUNCTS**

Figure 2: Hierarchy of verbal complements with focality preference

```
(3) <P id="p35" topic="t35.1">
  <S id="s35.1">
    <de ID="n173" firstm="MYES" syn-type="PR">
      <link Ltype="ident" POINT-BACK="n172"/>
      <W id="w35.1.1" lemma="il" pos="DET:def">La</W>
      <W id="w35.1.2" lemma="acqua" pos="NOM">Acqua</W>
      <W id="w35.1.3" lemma="marcio" pos="ADJ">Marcia</W></de>
      <W id="w35.1.4" lemma="potere" pos="VER:pres">può</W>
      <W id="w35.1.5" lemma="evitare" pos="VER:infi">evitare</W>
    <de ID="n521" firstm="MYES" syn-type="DefN">
      <W id="w35.1.6" lemma="il" pos="DET:def">il</W>
      <W id="w35.1.7" lemma="fallimento" pos="NOM">fallimento</W></de>
      <W id="w35.1.8" lemma="." pos="SENT">.</W></S>
  <S id="s35.2">
    <de ID="n174" firstm="MNO" syn-type="DefN-anch">
      <link Ltype="ident" POINT-BACK="n173"/>
      <W id="w35.2.1" lemma="il" pos="DET:def">La</W>
      <W id="w35.2.2" lemma="finanziaria" pos="NOM">finanziaria</W>
      <W id="w35.2.3" lemma="di" pos="PRE">di</W>
    <de ID="n522" syn-type="PR">
      <W id="w35.2.4" lemma="Vincenzo" pos="NPR">Vincenzo</W>
      <W id="w35.2.5" lemma="romagnoli" pos="NPR">Romagnoli</W></de>
  </de>... </S>...
</P>
```

Continue: Zero> Pronoun>clitic> Dem. N
 Retain: Clitic>Pronoun > Proper Name > Def. N >Def. N anchored> Zero > Dem. N
 Smooth Shift: Proper Name > Def. N > Pronoun>Def. N anchored
 Rough Shift: Def. N > Def. N anchored> Proper Name> Dem. N >Pronoun
 NULL: Proper name > Def. N anchored > Indef. N > Def. N

Figure 3: Transition types and referring expressions in Italian

Continue: Pronoun>Name>Def. N anchored
 Retain: Pronoun > Proper Name > Def. N anchored >Def. N>
 Smooth Shift: Proper Name > Def. N anchored > Pronoun>Def N
 Rough Shift: Def. N anchored> Proper Name> Def. N>Pronoun
 NULL: Proper name > Def. N anchored > Indef. N > Def. N

Figure 4: Transition types and referring expressions in Danish