On the Dimensions of Discourse Salience

Christian Chiarcos
Universität Potsdam

Abstract
This paper describes results of two corpus studies of information packaging of discourse referents in German dedicated to the following questions:

- Do sentence-initial position, pronominalization and subject role assignment reflect a single underlying dimension of discourse salience or multiple dimensions ?
- If there are multiple dimensions of salience, is it possible to associate them with a forward-looking and a backward-looking perspective on discourse, as proposed, e.g., in the context of Centering (Grosz et al., 1995) ?

This paper presents empirical findings from TüBa-D/Z, a corpus of German newspaper articles, that provide evidence against a unidimensional model of discourse salience, and support the claim that (at least) two dimensions of salience are to be distinguished and that these dimensions are associated with different temporal orientations on discourse.

1 Salience and Information Packaging

In the last 30 years, the notion of "salience" has been employed in many accounts for the information packaging of discourse referents, especially with respect to the choice of referring expressions (e.g., realization as definite NP or as a pronoun), and with respect to the pragmatic function of grammatical roles and word order preferences: Personal pronouns are assumed to represent more salient referents than nominals (Sgall et al., 1986; Ariel, 1990; Grosz et al., 1995), the left periphery of sentences (and in particular, the sentence-initial position) are associated with a high degree of salience (Sgall et al., 1986; Sridhar, 1988; Rambow, 1993), and the grammatical subject is assumed to serve a similar function (Fillmore, 1977; Tomlin, 1995; Grosz et al., 1995).

Consider the German example sentence in (1). The expression *sie* 'they' is a subject pronoun in preverbal (*vorfeld*) position; according to the aforementioned theories, its referent is to be regarded as highly salient, whereas the nominals *auf einem Tandem* 'on a tandem' and *ins Stadion* 'to the stadion' are postverbal non-subjects and thus non-salient.

(1) Sie wollen auf einem Tandem ins Stadion radeln
they want.to on a tandem into.the stadion go.by.bike
'They want to go to the stadion by tandem.' (TüBa-D/Z, sentence 113)

Despite the apparent agreement on the relevance of salience to different information packaging phenomena, researchers disagree on determinants and the actual nature of salience. Already Sridhar (1988, p.38) noted that 'a number of different factors have been claimed to contribute to salience', that '[r]esearchers are (...) divided on the effects

Proceedings of the Workshop "Beyond Semantics: Corpus-based Investigations of Pragmatic and Discourse Phenomena",
pages 31-44, Göttingen, Germany, 23-25 February 2011. Bochumer Linguistische Arbeitsberichte (3).

of salience to sentences', and further that 'salience is obviously (...) characterized by a number of superficially dissimilar properties'.

Since then, three major views on the nature of salience have been established:

unidimensional: In traditional unidimensional models as advocated by Sgall et al. (1986, word order and referring expressions), Gundel et al. (1993) and Ariel (1990, referring expressions), and Tomlin (1995, word order and grammatical roles), salience is seen as **a single dimension of cognitive states**: Every referent is assigned a particular degree of salience and this degree of salience determines the packaging preferences for this referent.

multifactorial: The radical antithesis to the unidimensional view is to abandon the idea of a generalized notion of salience, and to focus on the study of **individual factors**. This has been the premise of the psycholinguistic research of Osgood and Bock (1977) and Sridhar (1988), and has been recently revived by Kaiser and Trueswell (2004, to appear 2011) and Brown-Schmidt et al. (2005).

multidimensional: Multidimensional models of salience postulate the existence of multiple dimensions of salience as independent generalizations over a certain range of different factors. Typically, two dimensions are distinguished (Givón, 1983; Pattabhiraman and Cercone, 1990; Clamons et al., 1993; Mulkern, 2007): One dimension that is primarily defined with respect to the preceding discourse and/or the common ground, and that is thus primarily **backward-looking**. The other dimension is more concerned with the intentions and goals of the speaker and takes into consideration how these are manifested in subsequent discourse, and that is thus (at least partially) **forward-looking**.[1]

Psycholinguistic experiments and corpus studies on personal pronouns and demonstrative pronouns conducted by Kaiser and Trueswell (2004, to appear 2011), Brown-Schmidt et al. (2005) and Ellert and Hopp (2010) indicate that personal pronouns and demonstrative pronouns deviate in their antecedent selection preferences (in Finnish, Estonian, Dutch, English and German). In other words, certain salience factors contribute *independently* to the choice and interpretation of referring expressions in these languages. This observation can be seen as direct counterevidence for a unidimensional model that would postulate that demonstrative pronouns and personal pronouns reflect cognitive states organized in one uniform dimension of salience.

However, it is not necessary to conclude that multiple *cognitive* dimensions are involved: A functionalist with a unidimensional salience model might argue that the specific preference of personal pronouns to take subject antecedents can be attributed

[1] A representative multidimensional model of salience is the proposal of Clamons et al. (1993) and Mulkern (2007). They postulate that every referent is characterized by the degree of salience arising from the preceding context ('givenness', 'inherent salience') on the one hand, and on the other hand by the degree of salience imposed on this referent by the speaker ('importance', 'emphasis', 'imposed salience') in order to increase its accessibility in subsequent discourse. Similar approaches (with different terminologies) have been described by Givón (1983, 2001); Pattabhiraman and Cercone (1990), and are also suggested by Levelt (1989) and Chafe (1994).

to grammaticalization tendencies (comparable to those that lead to the development of syntactically bound (e.g., relative) pronouns out of anaphoric demonstrative pronouns in German, see Diessel, 1999, p.120ff), and further that these grammaticalization tendencies are actually based on conventional patterns of salience.[2] From a functional point of view, differences in subject-sensitivity are thus no counterevidence to a unidimensional salience model, because it is not only salience that affects packaging preferences but also grammatical conventions (that rely on the same conception of salience). This line of argumentation may be applied to every attempt to prove the insufficiency of unidimensional models of salience by showing that different types of referring expressions (or grammatical roles etc.) differ in their sensitivity to specific salience factors, especially those that address the linguistic realization of the antecedent.

In order to distangle grammaticalization tendencies from salience, it is thus necessary to evaluate predictions of unidimensional models of salience independently from the study of individual salience factors. This is the aim of the corpus study described in Sec. 3: Independently from the salience factors involved, a unidimensional model of salience predicts correlations between preverbal word order, pronominalization and subject role assignment. If the expected correlation between these phenomena cannot be confirmed, we have to conclude that at least two different dimensions or factors must be involved in the information packaging of these phenomena.

As an alternative to unidimensional models, Kaiser and Trueswell (2004, to appear 2011) suggest a multifactorial approach. A factor-based model, however, misses an important generalization, i.e., a theoretically motivated explanation for the underlying processes involved in information packaging, cf. the early critical remarks by Tomlin (1995). From a theoretical point of view – but also from the perspective of natural language processing (NLP) applications that try to interpret and reproduce information packaging preferences –, it is thus desirable to abstract from individual factors. Multidimensional models of salience provide such an abstraction in that they propose a dichotomy of logically independent dimensions of salience that interact in the process of information packaging. The second corpus study (Sec. 4) addresses the question whether these dimensions of salience correlate with forward-looking and backward-looking functions of referring expressions in discourse.

2 Corpus and Feature Extraction

The corpus studies described below are conducted on non-coordinated main clauses from the TüBa-D/Z corpus (v.5), a corpus of 2,213 German newspaper articles annotated for morphology, syntax and coreference (Telljohann et al., 2009; Naumann,

[2] In a unidimensional model of salience, grammatical roles and pronominalization are indirectly associated through the conception of salience: The subject represents the most salient referent of the current clause, and if we assume that this referent is most likely to remain the most salient referent of the following clause, it is to be expected to be realized as a personal pronoun then. By grammaticalization, the indirect association between subject role of the antecedent and the choice of a personal pronoun (originally mediated by the concept of salience) develops into a direct association, i.e., a grammatical convention that the original personal pronoun takes a subject antecedent.

2007). TüBa-D/Z is particularly well-suited for this study, as it combines anaphoric annotations with explicit annotations of topological fields of German sentences. With respect to word order, we can thus make use of the theoretically well-founded concept of *vorfeld* constituents.

From the corpus, all non-coordinated, non-embedded main clauses (40,713 clauses) were extracted, and all their nominal and pronominal arguments and adjuncts[3] were considered as (potential) referring expressions (79,222 in total).

The following classes of referring expressions (`ref`) were distinguished:

- 6 types of pronominal expressions
  - `perspron` personal pronoun, `pronadv` pronominal adverb, `dempron` demonstrative pronoun, `reflpron` reflexive pronoun, `pron` other pronouns (e.g., pronominal quantifiers)
- 7 types of nominal expressions
  - `name` proper name, `defNP` definite NP, `indefNP` indefinite NP[4], `possNP` possessive NP, `demNP` demonstrative NP, `NP` other NPs (e.g., NPs with semidemonstrative determiner *solch* 'such', or interrogative determiner *welch* 'which')
- coordinations
  - `coord,pron` pronominal coordination (all conjuncts are pronominal), `coord,NP` nominal coordination (at least one conjunct is nominal)

Depending of the parent nodes of the expressions under consideration in the syntax annotation, four different word order possibilities (`wo`) were distinguished.

- `vf` *vorfeld* positioning, preverbal (node label `VF`),
- `mf_initial` *mittelfeld* initial, immediately after the finite verb (NP/PP that is the *left-most* child of an `MF` node),
- `mf_noninitial` in the *mittelfeld*, but preceded by another expression,
- `nf` *nachfeld*, a right-peripheral field, following displaced verbal particles and infinite verbs (node label `NF`).

Three classes of grammatical roles (`gr`) were distinguished:

- `sbj`, grammatical subject (edge labels `on`, `onk`)
- `obj`, non-prepositional object (edge labels `oa`, `oak`, `od`, `odk`, `og`, `ogk`)
- `other`, remaining complements (incl. prepositional objects, other PPs, predications, etc.)

The values of `ref`, `wo` and `gr` represent the packaging phenomena distinguished in the first corpus study. For the second corpus study, two additional features, `given` and `important`, were derived from the coreference annotation:

- `given`, if linked to another expression in the preceding discourse by `coreferential`, `anaphoric`, `bound`, `cataphoric` or `instance` relations.
- `important`, if linked to another expression in the subsequent discourse by `coreferential`, `anaphoric`, `bound`, `cataphoric` or `instance` relations.

[3] `NX` and `PX` nodes directly attached to `VF`, `MF`, or `NF`

[4] As defined here, indefinite also includes determinerless and quantified NPs.

Feature extraction was performed using an extension of Gerlof Bouma's Prolog interface to TüBa-D/Z (Bouma, 2010).[5]

3 Predictions of the Unidimensional Model

The first corpus study investigates the predictions of the unidimensionality hypothesis with respect to expected correlations between *vorfeld* positioning (`vf`), subject role assignment (`sbj`) and pronominalisation (`perspron`) of referring expressions in German main clauses.

If *vorfeld* positioning, subject role assignment and pronominalisation all serve as indicators of a particularly high degree of a single dimension of salience, then `sbj` entails a particularly highly salient referent, this referent is thus to be represented as `perspron` (with respect to referring expressions) and `vf` (with respect to word order). In terms of conditioned probabilities, a unidimensional model of salience entails the following inequations:

$$P(\texttt{perspron}|\texttt{sbj}) \;>\; P(\texttt{perspron}) \tag{1}$$

$$P(\texttt{vf}|\texttt{sbj}) \;>\; P(\texttt{vf}) \tag{2}$$

or, more generally, for any two grammatical devices $X^{sal\uparrow}$ and $Y^{sal\uparrow}$ that are associated with particularly high degrees of salience:

$$P(X^{sal\uparrow}|Y^{sal\uparrow}) \;>\; P(X^{sal\uparrow}) \tag{3}$$

This formulation of the unidimensional model relies on the following additional assumptions:

(1) Pronominalization, subject role assignment and placement in the *vorfeld* are determined by grammatical (syntactic/semantic) and functional determinants.

(2) There is no semantic or syntactic constraint that disencourages the cooccurence of pronominalization, subject role and *vorfeld* positioning.

(3) Salience is the primary functional determinant of pronominalization, subject role and *vorfeld* positioning.

(4) Pronominalization, subject role and *vorfeld* positioning indicate high degrees of salience.

Assumption (1) states that the impact of language-independent factors on information packaging, e.g., biological factors, is marginal as compared to the impact of functional and grammatical factors. This is the fundamental assumption underlying the existing

[5]The code developed for this purpose is available under `http://www.ling.uni-potsdam.de/~chiarcos/` under "Resources".

literature on salience in discourse. Assumption (2) expresses the fact that a sentence as in ex. (1) is both syntactically well-formed and semantically felicitous. Assumption (3) is an assumption of any salience-based account of information packaging; (4) represents generally accepted claims on the impact of salience on information packaging.

Under these assumptions, a unidimensional model of salience predicts correlation between pronominalization, subject role assignent and *vorfeld* positioning, as the assumptions (1) to (4) entail that causal relationships between these packaging phenomena that are not mediated by salience are marginal if not inexistent.

As stated in assumptions (1) and (3), information packaging is, however, not exclusively determined by salience, but other factors may also play a role, although to a lower degree than salience: Subject role assignment is influenced, for example, by animacy. Also, word order preferences and pronominalization are affected by other factors besides salience. With one underlying dimension of salience, however, the effects of such circumstantial factors can be minimized if only those referents are considered that are marked as being salient with respect to *two* dimensions of information packaging, e.g., a referent that is both subject and in *vorfeld*. The expected minimization of such circumstantial factors can be captured in the following inequations:

$$P(\texttt{perspron}|\texttt{sbj},\texttt{vf}) \;\geq\; P(\texttt{perspron}|\texttt{sbj}) \tag{4}$$

$$P(\texttt{perspron}|\texttt{sbj},\texttt{vf}) \;\geq\; P(\texttt{perspron}|\texttt{vf}) \tag{5}$$

or, more generally

$$P(X^{sal\uparrow}|Y^{sal\uparrow}, Z^{sal\uparrow}) \;\geq\; P(X^{sal\uparrow}|Y^{sal\uparrow}) \tag{6}$$

Inequations (3) and (6) represent predictions that immediately follow from a unidimensional model of salience under the assumptions given above. If they cannot be confirmed in the data, then either one of the packaging phenomena under consideration is not actually a salience-indicating grammatical device (despite the support from the literature), or a unidimensional model of salience is inappropriate for the formalization of information packaging for the choice of referring expressions, the assignment of grammatical roles and word order preferences at the same time.

Table 1 summarizes the results obtained for inequation (3). For all grammatical devices, we can observe an increase of relative frequency under the condition of another salience-marking grammatical device as entailed by the unidimensionality hypothesis. This indicates that there is indeed a functional overlap between `perspron`, `sbj` and `vf`.

However, the marginal increase of `perspron` probability under the condition `vf` (and vice versa) may rise suspicions that the functional overlap between these three

realization $X^{sal\uparrow}$	condition $Y^{sal\uparrow}$	(conditioned) probability $P(X^{sal\uparrow}\|Y^{sal\uparrow})$	probability increase (vs. unconditioned)
perspron	(none)	10.80% (8,557/79,222)	
	vf	11.43%	+0.63%
	sbj	20.06%	+9.26%
sbj	(none)	42.50% (33,667/79,222)	
	perspron	78.94%	+36.44%
	vf	63.91%	+21.41%
vf	(none)	33.16% (16,789/79,222)	
	perspron	35.08%	+1.92%
	sbj	49.87%	+16.71%

Table 1: $P(X^{sal\uparrow}|Y^{sal\uparrow}) > P(X^{sal\uparrow})$ in TüBa-D/Z ?

realization		χ^2	ϕ
±perspron	±vf	p < .0001	.014
±perspron	±sbj	p < .0001	.257
±sbj	±vf	p < .0001	.305

Table 2: Significance (χ^2) and Pearson correlation coefficient (ϕ) of perspron, sbj, and vf

phenomena is actually a functional overlap between sbj and perspron on the one hand and between sbj and vf on the other hand, while vf and perspron are only loosely related. Nevertheless, also the latter correlation is highly significant and positive for all pairs of packaging phenomena considered, as shown in Table 2.[6]

While the corpus data does not contradict the predictions of (3), inequation (6) could not be confirmed: Against the expected increase of probability under the condition of two salience-marking grammatical devices as compared to a single salience-marking grammatical device, Table 3 shows a **decrease** of probability for subject pronouns in *vorfeld* (unlike pronouns under the condition of being subject, and *vorfeld* under the condition of being subject):

$$P(\texttt{perspron}|\texttt{vf},\texttt{sbj}) < P(\texttt{perspron}|\texttt{sbj}) \quad (7)$$

$$P(\texttt{vf}|\texttt{perspron},\texttt{sbj}) < P(\texttt{vf}|\texttt{sbj}) \quad (8)$$

Apparently, there is a bias against subject pronouns in *vorfeld* (albeit there is no evidence for a bias against pronouns **or** subjects in *vorfeld*). This is a clear violation of predictions of the unidimensionality hypothesis. As we excluded circumstantial factors and grammatical well-formedness conditions as potential causes for divergency, we have to conclude that there are (at least) two functional dimensions underlying pronominalization, subject role assignment and *vorfeld* positioning, and further, that

[6] In Table 2 and in the remainder of this paper, ±X means that X applies (e.g., +sbj) or that X does not apply (e.g., −sbj that matches obj and other).

realization	conditions		probability	probability increase vs.	
$X^{sal\uparrow}$	$Y^{sal\uparrow}$	$Z^{sal\uparrow}$	$P(X^{sal\uparrow}\|Y^{sal\uparrow}, Z^{sal\uparrow})$	$P(X^{sal\uparrow}\|Y^{sal\uparrow})$	$P(X^{sal\uparrow}\|Z^{sal\uparrow})$
perspron	vf	sbj	15.51% (2,604/16,789)	+4.08%	**−4.55%**
vf	perspron	sbj	38.55% (2,604/6,755)	+3.47%	**−11.32%**
sbj	vf	perspron	86.74% (2,604/3,002)	+22.84%	+7.80%

Table 3: $P(X^{sal\uparrow}|Y^{sal\uparrow}, Z^{sal\uparrow}) \geq P(X^{sal\uparrow}|Y^{sal\uparrow})$ in TüBa-D/Z ?

subject role assignment is associated with both dimensions.

As for contextual features involved with these dimensions, the structure of the preceding discourse is generally assumed to play an important role: Previous mention and the linguistic realization of the antecedent are commonly considered to be a major determinant of pronominalization (Sgall et al., 1986; Ariel, 1990; Grosz et al., 1995), it is assumed to be associated with subject role assignment (Prince, 1992; Lambrecht, 1994, also cf. preference for continue transitions in Grosz et al., 1995), and traditionally with *vorfeld* positioning as well (the original working hypotheses of Speyer, 2007, and Dipper and Zinsmeister, 2009). A number of recent corpus studies, however, could not confirm that the preceding discourse determines *vorfeld* positioning, and a number of alternative factors have been suggested in consequence:

- Evidence against the primarily anaphoric nature of the *vorfeld* can be drawn from a number of recent corpus studies that actually attempted to *prove* the relevance of the preceding context to *vorfeld* positioning: For a collection of German prose text from various genres, Speyer (2007) reported that 51% of *vorfeld* constituents could be neither semantically nor anaphorically linked to the preceding discourse. On a corpus of parliamentary debates, Dipper and Zinsmeister (2009) found that 55% of *vorfeld* constituents stand in no obvious relationship to the preceding discourse, whereas only 23% are anaphorically linked. In their study of object arguments in the *vorfeld* of main clauses in the NEGRA corpus, Weber and Müller (2004) found no indication that anaphoric (given/definite/pronominal) objects tend to precede non-anaphoric (new/indefinite/nominal) subjects. In fact, indefinite objects preceded definite subjects in OVS sentences more often than vice versa.
- A smaller number of corpus studies and theoretical papers have dealt with alternative factors contributing to *vorfeld* positioning: For example, Filippova and Strube (2007) found that *vorfeld* constituents tend to refer to the global discourse topic (i.e., names mentioned in the headlines in their collection of biographic articles). Speyer (2007) proposed that the *vorfeld* is the preferred locus of contrastive ex-

pressions and frame-setting topics, whereas purely anaphoric expressions are positioned there only if the *vorfeld* would have been left unoccupied otherwise (cf. Frey (2004a) for a similar model).

Previously proposed non-anaphoric factors of *vorfeld* positioning often involve certain intentions on the side of the speaker, i.e., to express contrastivity, importance or to make sure that subsequent of information are interpreted in the context of a particular situational environment. In the words of Lötscher (1984), these functions may be seen as specific aspects of the 'highlighting' function of the *vorfeld*.[7]

Of course, it is problematic to quantify 'highlighting' without having direct access to the mental discourse model of the speaker at the moment of uttering. But at least one aspect of 'highlighting' can be extrapolated from the text itself – the speaker's intention to prepare the hearer for the forthcoming discourse: By placing the referent in preverbal position (or by choosing an otherwise prominent realization), the speaker performs a 'foregrounding' operation whose effects on the subsequent discourse can be observed, in particular, the increased anaphoric accessibility of the referent. The speaker's foregrounding intentions can thus be inferred from the distribution and the realization of the referent in the forthcoming discourse. As far as foregrounding is concerned, 'highlighting' can thus be approximated by forward-looking factors.[8]

The second corpus study evaluates the hypothesis that the dimensions of salience involved in *vorfeld* positioning, subject role assignment and pronominalisation can be aligned with such a backward-looking/forward-looking dichotomy.

4 Temporal Dimensions of Salience

The dichotomy between forward-looking and backward-looking salience is adopted in most multidimensional models of salience (Givón, 1983, 2001; Clamons et al., 1993; Mulkern, 2007, see also Grosz et al., 1995), and the corpus study described in this section investigates the relevance of this distinction for the packaging phenomena under investigation here.

As maximally theory-independent metrics of salience, backward-looking salience is reduced here to the existence of a previous reference to the same referent (`+given`), forward-looking salience is approximated by the existence of a subsequent mention of the same referent (`+important`).

A series of χ^2 square tests where the features `±perspron`, `±sbj` and `±vf` were tested against `±given` and `±important` reveals a significant interaction and a pos-

[7] Alternative terms include 'newsworthiness' (Mithun, 1992), 'imposed salience' (Clamons et al., 1993; Mulkern, 2007), or 'importance' (Givón, 1988), all discussed with respect to word order inverse and fronting in multiple languages.

[8] While this does not mean that forward-looking salience can be **equated** with the speaker's intention to highlight certain referents, forward-looking factors allow to reconstruct a certain fraction of the speaker's intentions at the moment of utterance, but only those that deal with his intention to prepare the hearer for the development of the subsequent discourse in order to guide him to a specific insight. That conversation involves such antipacitory elements was already emphasized by Grosz et al. (1995) who mention that speakers 'should plan ahead to minimize the number of shifts'. Of course, other intentions of the speaker, e.g., emotions or the intention to trigger certain implicatures (Ariel, 1990; Gundel et al., 1993) cannot be reconstructed with this heuristic. Every approximation of the speaker's original intentions by means of forward-looking factors is thus incomplete, but nevertheless feasible with respect to the foregrounding function of grammatical devices.

realization	±given χ^2	±given ϕ	±important χ^2	±important ϕ
±perspron	$p < .0001$	.342	$p < .0001$	.174
±sbj	$p < .0001$	.288	$p < .0001$	.279
±vf	$p < .0001$	.065	$p < .0001$	.073

Table 4: Significance (χ^2) and Pearson correlation coefficient (ϕ) of ±given/ ±important and packaging phenomena

itive correlation between the packaging phenomena considered and both dimensions of salience (Table 4).

In order to assess how ±given and ±important interact during the derivation of packaging preferences, C4.5 decision trees were trained on the feature sets extracted from TüBa-D/Z (using the J48 implementation of WEKA, Witten and Frank, 2005): The C4.5 algorithm maximizes the correctness of classification, and with ±given and ±important as input features and different packaging phenomena as target classification, the decision tree built up by algorithm allows to extrapolate the influence of previous mention and of subsequent mention on the choice of referring expressions, the assignment of grammatical roles and word order preferences.

(a) referring expressions (ref)

```
correctness: 34.6%
(baseline: defNP, 33.6%)

+given
|   +important: perspron
|   -important: defNP
-given: defNP
```

(b) grammatical roles (gr)

```
correctness: 53.1%
(baseline: sbj, 42.5%)

+given: sbj
-given
|   +important: sbj
|   -important: other
```

(c) word order (wo)

```
correctness: 38.7%
(baseline: mf_initial, 33.6%)

+given: mf_initial
-given
|   +important: vf
|   -important: mf_noninitial
```

Figure 1: C4.5 decision trees to predict packaging preferences from ±given and ±important.

For every dimension of information considered here, ref, gr and wo, a decision tree was trained to predict the actual grammatical device (with the fine-grained subclasses as described in Sec. 2) based on the features ±given and ±important. The resulting trees are shown in Fig. 1. Compared with the baseline (most frequent class), all classifiers yield an increase in correctness.[9] A more interesting evaluation method is a comparison of the classifier strategies with claims in linguistic literature:

[9]Note that these classifiers only serve as an indicator of the way that ±given and ±important influence information packaging. The overall classification results are poor, but mostly because the number of packaging phenomena distinguished for the different levels of information packaging is far greater than the number of possible combination of input values. However, with more fine-grained measurements of backward-looking and forward-looking salience, as studied, for example, by Chiarcos (2009), more detailed information packaging predictions may be possible. At this point, such improvements are left as a topic for subsequent research.

We can observe a remarkable degree of compatibility of the classifiers with theoretical models.

- As expected from the literature, the `ref` classifier predicts a close association between `given` and `perspron`. That `important` has an impact on pronominalization may reflect a preference to maintain an established (pronominal) topic over several utterances, cf. Lambrecht (1994, p.199ff.), Grosz et al. (1995).
- The `gr` classifier combines two conflicting views on functional determinants of subjecthood found in two different branches in the literature: Traditionally, the subject is associated with high degrees of backward-looking salience (e.g., Prince, 1992), but in typological literature, it is assumed that subjects serve an attention-guiding, foregrounding function (Tomlin, 1995; Pustet, 1997).[10] The classifier combines both views by stating that the subject is important *or* given.
- The `wo` classifier closely resembles modern approaches on *vorfeld* positioning in German: Frey (2004a,b) postulated that the unmarked position of the (given-ness-) topic is the immediate post-verbal position (the preferred locus of *given* referents according to the `wo` classifier), whereas placement of the topic in the *vorfeld* requires the presence of an another pragmatic force, e.g., 'kontrast' as defined by Vallduví and Vilkuna (1998). Above, a functional resemblance between foregrounding and contrast was suggested (both represent different aspects of the 'highlighting' force of the *vorfeld*), so that the preference to place subsequently mentioned referents in `vf` can be compared to the effect of kontrast in Frey's model.

  The predicted effects of the feature bundles `+given` and `−given/+important` on `wo` can also be compared to Lambrecht's information-structural characterization of the left periphery: Lambrecht (1994, p.199ff) assumes that the sentence-initial position serves the function of *topic announcement*, i.e., that a referent that has not been established as a topic before (`−given`) is marked as being the topic of the subsequent discourse segment (`+important`). As opposed to this, the function of *topic maintenance* of already established (`+given`) topics is not associated with the left periphery, but with proximity to the finite verb, i.e., `mf_initial` in German main clauses.

As mentioned above, decision trees can be rephrased as rules for information packaging preferences and thus compared with regularities reported in the linguistic literature. Table 5 summarizes these rules and reveals another remarkable coincidence: `±given` and `±important` predict exactly the distribution of grammatical devices observed in the first corpus study:

[10] One of the few models that combines both views can be found in Centering (Grosz et al., 1995) where subjects are ascribed both a forward-looking function (the subject is the preferred center, i.e., highest-ranking forward-looking center), and a backward-looking function (preference for identity of preferred center and backward-looking center, preference of continuity of the backward-looking center).

±given	±important	prediction		
+	+	perspron	sbj	mf_initial
+	−	defNP	sbj	mf_initial
−	+	defNP	sbj	vf
−	−	defNP	other	mf_noninitial

Table 5: Information packaging preferences predicted from ±given and ±important

(a) an association between pronominalization and subject role (+given, +important),

(b) an association between *vorfeld* positioning and subject role (−given, +important), and

(c) a dispreference for subject pronouns (+given) to coincide with *vorfeld* (−given).

5 Results

Taken together, both corpus studies provide evidence against a unidimensional model of discourse salience, and, more specifically, they support the claim that (at least) two dimensions of discourse salience are to be distinguished, and further that these dimensions are associated with different temporal orientations on discourse:

- It is necessary to distinguish at least two dimensions of salience in order to account for *vorfeld* positioning, pronominalization and subject role assignment in a salience-based model.
- Previous mention (backward-looking salience, givenness) and subsequent mention (forward-looking salience, importance) have a highly significant influence on these packaging phenomena.
- The interaction between both factors leads to the observed distribution of these packaging phenomena.

Acknowledgements

The research described in this paper was was funded by the German Research Foundation (DFG) in the context of the Collaborative Research Center (SFB) 632, Project D1 "Linguistic Database" at the Universität Potsdam, Germany. I would like to thank three anonymous reviewers, Stefanie Dipper, Heike Zinsmeister, Julia Ritz and Robin Hörnig for comments and feedback, and Gerlof Bouma for his support during the development of the feature extraction scripts.

References

Mira Ariel. *Accessing Noun-Phrase Antecedents*. Routledge, London, New York, 1990.

Gerlof Bouma. Syntactic tree queries in Prolog. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV), held in conjunction with ACL 2010*, pages 212–216, Uppsala, Sweden, July 2010.

Sarah Brown-Schmidt, Donna K. Byron, and Michael K. Tanenhaus. Beyond salience: Interpretation of personal and demonstrative pronouns. *Journal of Memory and Language*, 53:292–313, 2005.

Wallace Chafe. *Discourse, Consiousness, and Time. The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press, Chicago and London, 1994.

Christian Chiarcos. *Mental Salience and Grammatical Form. Toward a Model of Salience for Natural Language Generation*. PhD thesis, Universität Potsdam, Germany, Nov 2009.

C. Robin Clamons, Ann E. Mulkern, and Gerald Sanders. Salience signaling in Oromo. *Journal of Pragmatics*, 19:519–536, 1993.

Holger Diessel. *Demonstratives. Form, Function, and Grammaticalization*. John Benjamins, Amsterdam, Philadelphia, 1999.

Stefanie Dipper and Heike Zinsmeister. The role of the German vorfeld for local coherence: A pilot study. In *Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL)*, pages 69–79, Narr, Tübingen, Sep 2009.

Miriam Ellert and Holger Hopp. Disentangling topicality from order of mention in the resolution of the German subject pronouns *er* and *der*: Off-line and on-line data. In *Proceedings of the Biennal Conference of the German Society of Cognitive Science (KogWis 2010)*, Potsdam, Germany, Oct 2010.

Katja Filippova and Michael Strube. The German vorfeld and local coherence. *Journal of Logic, Language and Information*, 16(4):465–485, 2007.

Charles J. Fillmore. Topics in lexical semantics. In Roger W. Cole, editor, *Current Issues in Linguistic Theory*, pages 76–138. Indiana University Press, Bloomington, 1977.

Werner Frey. The grammar-pragmatics interface and the German prefield. *Sprache und Pragmatik*, 52:1–39, 2004a.

Werner Frey. A medial topic position for German. *Linguistische Berichte*, 198:153–190, 2004b.

Talmy Givón. Introduction. In Talmy Givón, editor, *Topic Continuity in Discourse: A Quantitative Cross-Language Study*, pages 5–41. John Benjamins, Amsterdam and Philadelphia, 1983.

Talmy Givón. The pragmatics of word order: Predictability, importance and attention. In Michael Hammond, Edith A. Moravcsik, and Jessica Wirth, editors, *Studies in Syntactic Typology*, pages 243 – 284. John Benjamins, Amsterdam and Philadelphia, 1988.

Talmy Givón. *Syntax*. John Benjamins, Amsterdam and Philadelphia, 2001.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

Jeanette K. Gundel, Nancy A. Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):247–307, 1993.

Elsi Kaiser and John Trueswell. The referential properties of Dutch pronouns and demonstratives: Is salience enough? In Matthias Weisgerber and Cecile Meier, editors, *Proceedings of Sinn und Bedeutung 8*, University of Konstanz linguistics working papers, pages 137–150, Konstanz, 2004.

Elsi Kaiser and John Trueswell. Investigating the interpretation of pronouns and demonstratives in Finnish: Going beyond salience. In Edward Gibson and Neal J. Pearlmutter, editors, *The Processing and Acquisition of Reference*. MIT Press, Cambridge, Mass, to appear 2011.

Knud Lambrecht. *Information Structure and Sentence Form. Topic, Focus and the Mental Representations of Discourse Referents*. Cambridge University Press, Cambridge, 1994.

Willem J.M. Levelt. *Speaking: From Intention to Articulation*. MIT Press, 1989.

Andreas Lötscher. Satzgliedstellung und funktionale Satzperspektive. In Gerhard Stickel, editor, *Pragmatik in der Grammatik*, Jahrbuch 1983 des Instituts für deutsche Sprache, pages 118–151. Schwann, Düsseldorf, 1984.

Marianne Mithun. Is basic word order universal? In Doris L. Payne, editor, *Pragmatics of Word Order Flexibility*, pages 15–62. John Benjamins, Amsterdam and Philadelphia, 1992.

Ann E. Mulkern. Knowing who's important: Relative discourse salience and Irish pronominal forms. In Nancy A. Hedberg and Ron Zacharski, editors, *The Grammar-Pragmatics Interface: Essays in honor of Jeanette K. Gundel*, pages 113–142. John Benjamins, Amsterdam and Philadelphia, 2007.

Karin Naumann. Manual for the Annotation of in-document Referential Relations. Technical report, Universität Tübingen, Seminar für Sprachwissenschaft, 2007. version of May 2007.

Charles E. Osgood and J. Kathryn Bock. Salience and sentencing: Some production principles. In Sheldon Rosenberg, editor, *Sentence Production: Developments in Research and Theory*, pages 89–140. Erlbaum, Hilldale, N.J., 1977.

T(hiyagarajasarma) Pattabhiraman and Nick Cercone. Selection: Salience, relevance and the coupling between domain-level tasks and text planning. In *Proceedings of the 5th International Workshop on Natural Language Generation (IWNLG 1990)*, pages 79–86, Pittsburgh, Apr 1990.

Ellen F. Prince. The ZPG letter: Subjects, definiteness, and information-status. In Sandra A. Thompson and William C. Mann, editors, *Discourse Description: Diverse Analyses of a Fund Raising Text*, pages 295–325. John Benjamins, Amsterdam and Philadelphia, 1992.

Regina Pustet. *Diskursprominenz und Rollensemantik – Eine funktionale Typologie von Partizipantensystemen*. Lincom Europa, München, 1997.

Owen Rambow. Pragmatic aspects of scrambling and topicalization in German. In *Workshop on Centering Theory in Naturally-Occurring Discourse*. Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, PA, 1993.

Petr Sgall, Eva Hajićová, and Jarmila Panevova. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. Reidel, Dordrecht, 1986.

Augustin Speyer. Die Bedeutung der Centering Theory für Fragen der Vorfeldbesetzung im Deutschen. *Zeitschrift für Sprachwissenschaft*, 26(1):83–116, 2007.

Shikaripur N. Sridhar. *Cognition and Sentence Production. A Cross-Linguistic Study*. Springer, New York and Berlin, 1988.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen, 2009. version of November 2009.

Russel S. Tomlin. Focal attention, voice, and word order. An experimental, cross-linguistic study. In Mickey Noonan and Pamela Downing, editors, *Word Order in Discourse*, pages 517–554. John Benjamins, Amsterdam and Philadelphia, 1995.

Enric Vallduví and Maria Vilkuna. On rheme and kontrast. In Peter Culicover and Louise McNally, editors, *The Limits of Syntax*, pages 79–108. Academic Press, New York, 1998.

Andrea Weber and Karin Müller. Word order variation in German main clauses: A corpus analysis. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora. Held in Conjunction with the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 71–78, Geneva, August 2004.

Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Pub, 2005.