Evaluating Salience Metrics for the Context-Adequate Realization of Discourse Referents

Christian Chiarcos
chiarcos@uni-potsdam.de
Universität Potsdam

Abstract
We describe the application of a framework for salience metrics and linguistic variability with respect to the contextually adequate choice of referring expressions and grammatical roles: Where multiple meaning-equivalent candidate realizations are available that differ in one of these aspects, NLG systems can apply salience metrics to predict contextually adequate realization preferences. We evaluate this claim and a number of parameters of salience metrics found in the theoretical literature on two German newspaper corpora. Key features of the approach described here include the application of a two-dimensional model of salience, how its theoretical predictions can be exploited to develop salience metrics for a particular phenomenon, and that these salience metrics can be subsequently applied to other phenomena. This approach can be applied to develop classifiers to predict packaging preferences for phenomena where little training data is available.

1 Motivation and Background
For an example sentence from the RST Discourse Treebank (Carlson et al., 2003, file 3), example (1) illustrates how the same ‘thought’ can be realized, or ‘packaged’ (Chafe, 1976) in many different ways: Three referents, the insurance agent Toni, her sister Cynthia and their apartment suffer from an earthquake, the central protagonist of the paragraph is Toni, and the text goes on elaborating her situation.

(1) The apartment she shares with her sister was rattled ...
   (a) The apartment the agent shares with her sister ...
   (b) The earthquake rattled the apartment she shares ...

We consider two packaging phenomena: Referring expressions (1a: definite NP vs. pronoun), and grammatical roles (1b: active vs. passive).¹

These variants are meaning-equivalent in the sense of Dorr et al. (2004), but according to theories of referential coherence (Šgall et al., 1986; Grosz et al., 1995; Givón, 2001), they express different discourse functions, often described with reference to the notion of ‘discourse salience’.² Accordingly, the local discourse context – or, better, a salience score calculated on this basis – can help to predict contextually adequate packaging preferences.

In NLG, discourse salience has been employed to generate referring expressions (McCoy and Strube, 1999), to assign grammatical roles (Stede, 1998), and word order preferences (Kruijff et al., 2001). More recently, however, salience-based approaches have been increasingly superseded by statistical approaches, that nevertheless build on earlier theories of salience, e.g., Shiramatsu et al. (2007) for referring expressions, Zarrieß et al. (2011) for voice alternation, and Cahill and Riester (2009) for word order. One of the reasons for this methodological shift may be the observation (noted, for example, by

¹Along with referring expressions and grammatical roles, word order alternation has been described in a similar way, and it is of particular importance for the motivation of two-dimensional models of salience (Chiarcos, 2011b). For reasons of space, however, this paper concentrates on referring expressions and grammatical roles.

²Discourse salience is to be distinguished from other types of salience, that are either not specific to discourse referents (e.g., salience of semantic features, Ortony et al. 1985), or defined with respect to other modalities (e.g., visual salience, Itt 2003, Kelleher 2011).
Navaretta, 2002) that the existing approaches developed until the late 1990s were only partially compatible with each other, as they employed different theories of referential coherence.

Major theories of referential coherence, e.g., Centering (Grosz et al., 1995), its instantiations (Poesio et al., 2004), Topicality (Givón, 2001) and Functional Generative Description (Sgall et al., 1986, FGD) share a set of common insights, in particular, the close association between referential coherence and attentional states (as manifested in the salience of discourse referents), but they focus on different aspects of referential coherence and formalize them in different ways.\(^3\)

Even worse, the field is notoriously plagued by a multitude of incompatible terminologies: ‘Salience’, for example, is used as a near-synonym of ‘givenness’ (Sgall et al., 1986, p.54f.), but also as a near-synonym of ‘newness (for the hearer)’ (Davis and Hirschberg, 1988), or ‘degree of interest (of the speaker)’ (Langacker, 1997, p.22). Therefore, the operationalization of discourse salience in NLG requires a theoretical foundation and a formalization of salience and its effects on information packaging.

This paper takes its point of departure from a theoretical framework of discourse salience that has been developed as a generalization over Centering, Topicality and FGD. This framework, as sketched in Sect. 2, resolves the terminological difficulties associated with the notion of salience by distinguishing two dimensions of salience, with independent effects on referring expressions, grammatical roles and word order. One advantage of this theory-based approach as compared to a plain statistical classifier is that it incorporates a set of theoretical assumptions that guide the development of salience metrics, and that predict an impact of a salience metric even on phenomena not considered during the development of this particular metric.

Section 3 identifies a number of parameters that allow to reconstruct different instantiations of Centering, Topicality and FGD salience within this model. Section 4 deals with the empirical evaluation of these parameters on two German newspaper corpora, in Sections 4.1 and 4.2 elementary metrics for both dimensions of salience are developed, and Sect. 4.3 confirms theoretical predictions on the impact of both dimensions of salience on noun phrase complexity and grammatical roles.

2 A Framework of Salience in Discourse

Inspired by Givón’s topicality measurements and hierarchies of grammatical devices associated with them, Chiarcos (2010; 2011a) developed an operationalizable formalization of functional-cognitive theories of information packaging within the Mental Salience Framework (MSF), a framework for the development and interpretation of salience metrics in discourse. Below, we sketch the reconstruction of Centering, Topicality and FGD salience within this approach. We provide a brief, technical description only, as the focus of this paper is to evaluate the resulting salience metrics.

The framework, schematically illustrated in Fig. 1, consists of the following components:

- a theoretical model of salience, grounded in cognitive linguistics and functional grammar (Chiarcos, 2011a),
- the specification of two dimensions of salience, backward-looking hearer salience, and forward-looking speaker salience (Sect. 2.1), and the corresponding metrics (Sect. 2.2),
- packaging hierarchies, i.e., rankings of grammatical devices for different packaging phenomena (Sect. 2.3), that are aligned with accumulated salience scores calculated from hearer salience and speaker salience (Sect. 2.4), and
- principles for the mapping between packaging hierarchies and salience scores (Sect. 2.5).

As opposed to related models in functional-cognitive linguistics, e.g., Mulkern (2007), our formalization is operationalizable for NLG applications: It allows to predict packaging preferences for discourse referents from numerical salience scores (Sect. 2.5).

Metrics of salience applied in Natural Language Processing are dominated by research on anaphora

---

\(^3\)For example, Grosz et al.’s Centering posits an adjacency constraint, whereas FGD and Topicality employ distance measurements. FGD predicts constraints on word order and referring expressions, but it differs from Centering and Topicality in that in formalizes only the backward-looking aspect of salience in discourse.
resolution in the tradition of Lappin and Leass (1994). Such salience metrics do, however, focus on the backward-looking, hearer-oriented aspect of salience, whereas the speaker-oriented, forward-looking aspect of salience is neglected. This tradition also had a strong impact on NLG, in particular in the field of generating referring expressions (GRE). Current metrics of discourse salience in GRE are thus essentially concerned with hearer salience, although the relevance of speaker-oriented factors has been recognized for other aspects of NLG, e.g., for German word order as being sensitive to a domain-specific ‘aboutness’ criterion (Filippova and Strube, 2007).

Within the MSF, Centering, Topicality and FGD salience can be reconstructed as configurations of hearer and speaker salience. As opposed to earlier generalizations over some of these theories, e.g., Krahmer and Theune (2002), this paper adopts a two-dimensional model of salience for NLG. This bidimensionality not only helps to resolve conflicts between different terminological traditions, it also accounts for newer evidence that many packaging phenomena require the differentiation of (at least) two dimensions of discourse salience (Kaiser and Trueswell, 2010; Chiarcos, 2011b).

The most important parameters are summarized in Sect. 3.

2.1 Salience

In neurobiology and psychology, salience is defined as a gradual assessment of attentional states (Itti et al., 2005), and it is used in this sense also in functional grammar (Sgall et al., 1986), cognitive linguistics (Talmy, 2000) and computational linguistics (Grosz et al., 1995). In order to resolve the terminological difficulties mentioned above, we distinguish two dimensions of salience in discourse associated with different roles regarding the flow of attention in discourse.

From the perspective of an NLG system, ‘attentional states’ are primarily those of the speaker: Information that is relevant to the speaker is more salient than information not considered relevant (Pattabhiraman, 1992; Reed, 2002). Beyond this, a cooperative speaker takes the perspective of the addressee into consideration, i.e., she acts according to her assumptions about the attentional states of the hearer (Prince, 1981). Generating text that is both coherent (for the hearer) and goal-directed (for the speaker) requires both perspectives.

The resulting multidimensionality of salience is not specific to dialog, but has also been confirmed for written, monologuous discourse, e.g., by Kaiser and Trueswell (2010) and Chiarcos (2011b). The latter also provides evidence for a differentiation between a backward-looking and a forward-looking dimension of salience. Taking up Centering terminology, assumed attentional states of the hearer can indeed be characterized as being primarily backward-looking (the preceding discourse allows to approximate the attentional states of the hearer), whereas attentional states of the speaker involve a forward-looking aspect (subsequent discourse can unveil the speaker’s earlier intentions to elaborate on a particular issue).

This difference is modelled here by distinguishing two independent dimensions of discourse salience: (i) speaker salience represents the attentional states of the speaker (that express her intentions to guide the hearer’s focus of attention), and (ii) hearer salience represents the speaker’s approximation of the attentional states of the hearer.

Cross-linguistic research indicates that both aspects of attention control in discourse are necessary to choose of referring expressions, and to assign grammatical roles appropriately.\(^5\)

2.2 Salience metrics

Salience is represented by means of numerical scores, so that a principally unlimited number of attentional states can be distinguished, cf. Sgall et al. (1986), Ariel (1990), and Lappin and Leass

\(^4\)This is true even for multidimensional models of salience in GRE such as van der Sluis and Krahmer (2001): Their ‘focus-space salience’ is concerned with the visual environment, ‘inherent salience’ is a semantic criterion (uniqueness within a domain), ‘linguistic salience’ is the hearer-oriented, backward-looking aspect of discourse salience.

\(^5\)Referring expressions are associated with hearer salience (Ariel, 1990; Gundel et al., 1993; for German see Heusinger, 1997); demonstratives also with speaker salience (attention guidance, contrast) (Ehlich, 1982; Diessel, 2006; for German see Bosch et al., 2007). The assignment of grammatical roles is sensitive to hearer salience (Fillmore, 1977; Sgall et al., 1986) as well as speaker salience (foregrounding) (Pustet, 1995; Tomlin, 1995).
The salience of a referent \( r \) is assessed by means of one metric of hearer salience, \( \text{hsal}(r) \), and one metric of speaker salience, \( \text{ssal}(r) \). Backward-looking salience factors that pertain to the preceding discourse are available to both speaker and hearer; they represent primarily factors of hearer salience. Forward-looking factors that take the subsequent discourse into consideration are factors of speaker salience: If the speaker intended to guide the hearer’s attention in a planful way – to prepare him for the following development of discourse – the subsequent discourse provides a rough approximation of the speaker’s intentions at the moment the current utterance was produced.

For a referent \( r \), the salience factor \( i \) is represented as a numerical value \( x_{ri} \) with \( 0 \leq x_{ri} < 2 \). Hearer salience and speaker salience are calculated from the weighted sum of these factors. The weights \( w_{i,\text{hsal}} \in \mathbb{R} \) and \( w_{i,\text{ssal}} \in \mathbb{R} \) correspond to the relative impact that a particular salience factor \( x_{ri} \) has on the salience scores \( \text{hsal}(r) \) and \( \text{ssal}(r) \). If \( x_{ri} \) is speaker-private, then \( w_{j,\text{hsal}} = 0 \).

Salience scores are normalized to the range \( 0 \leq \text{sal}(r) < 2 \): Scores greater than 1 indicate a high degree of salience, 0 the absence of salience. For distance-sensitive factors of hearer salience, we employ the normalization function \( n(x, k) = \frac{x}{k+1} \) where \( k \) represents the distance from the last mention of the referent (e.g., the number of intermediate clauses), and \( x \) the salience score that the referent would have if the last mention was in the preceding utterance. All theories mentioned above assume that a referent \( r \) mentioned in the last utterance is more hearer salient than any referent in the utterance before, i.e., \( x > n(2, 1) = \frac{2}{3} \). We thus adopt 0.8 as minimum value for \( x \). For presentational reasons, we further assume that 1.0 is the average hearer salience score for a referent mentioned in the preceding utterance, possible values of \( x \) are thus normalized to the range \( 0.8 \leq x \leq 1.2 \).

### 2.3 Packaging hierarchies

Figures 2 and 3 illustrate the predicted impact of salience on referring expressions and grammatical roles. These hierarchies generalize over several theories mentioned above and assume that a referent’s hearer salience is calculated not only from its mention, but if salience scores from the entire referential chain are added up (as in Lappin and Leass’ original proposal). This paper, however, follows Centering, Topicality and FGD and only considers the last mention of the referent.

---

6Hearer salience scores greater than 1.2 are obtained if a referent’s hearer salience is calculated not only from its mention, but if salience scores from the entire referential chain are added up (as in Lappin and Leass’ original proposal). This paper, however, follows Centering, Topicality and FGD and only considers the last mention of the referent.
rankings and scales of grammatical devices developed in cognitive and functional linguistics (footnote 5): They are assumed to be applicable cross-linguistically, and also to English (Chafe, 1994; Cornish, 2007; Fillmore, 1977; Tomlin, 1995), and thus illustrated for ex. 1:

(1a): In accordance with Fig. 2, the use of the agent in place of the pronoun is possible as a means to express a high degree of speaker salience, e.g., in order to put Toni in the foreground. However, as Toni already is the maximally hearer salient referent in the preceding discourse, this is not necessary and thus avoided.

(1b): In the original, Toni is the subject of a relative clause attached to the subject apartment. In (1b), the relative clause is attached to the direct object, and in accordance with Fig. 3, this indicates a lower degree of hearer salience and speaker salience as compared to the original realization. This is justified only if the earthquake was speaker salient, e.g., because it would be the intended main protagonist of the following sentences (what it isn’t), (1b) is thus dispreferred as it would distract the hearer’s focus of attention from Toni.

2.4 Cumulated salience scores

We employ cumulated salience scores for the mapping between salience scores and packaging hierarchies: For every packaging phenomenon, the cumulated salience score is the weighted sum of hearer salience score $hsal(r)$ and speaker salience score $ssal(r)$, i.e., $ref(r)$ for referring expressions and $gr(r)$ for grammatical roles.

$$ref(r) := u_{hsal,ref} hsal(r) + u_{ssal,ref} ssal(r)$$
$$gr(r) := u_{hsal,gr} hsal(r) + u_{ssal,gr} ssal(r)$$

As a convention, the realization favored by a high degree of hearer salience is associated with high, positive cumulated salience scores. If a high degree of speaker salience favors the same realization, $ssal$ is assigned a positive weight (as for $gr(r)$), if $ssal$ favors a deviation from $hsal$ preferences, it is assigned a negative weight (as for $ref(r)$).

In practical application, the relative weights of $hsal$ and $ssal$ for a particular phenomenon, say, sentence-initial word order, can be trained with a simple Multi-Layer Perceptron (MLP) with one hidden node: $hsal$ and $ssal$ scores serve as input nodes and two nodes representing ± initial as output nodes. After training the MLP, the weights of hearer salience and speaker salience can be extrapolated from the activation function of the hidden node.

2.5 Predicting packaging preferences

Cumulated salience scores are interpreted against a packaging hierarchy by means of hierarchy alignment: The referent with the highest cumulated salience score is assigned the highest-ranking grammatical device available, etc. For grammatical roles, for example, the candidate realization would be preferred that minimizes the deviations between the salience ranking of discourse referents and their relative syntactic prominence (e.g., when a highly referent is assigned object role while a non-salient referent is assigned subject role).

This hierarchy alignment, as well as additional realization thresholds that express, for example, that pronouns require a certain minimum of salience, can be implemented as constraints in an optimality-theoretic setting. Alternatively, alignment between salience scores and their most likely realization can also be formulated as a minimization problem, so that standard approaches to optimization problems can be applied (Pattabhiraman, 1992). A similar ranking-based approach has been applied, for example, by Zarrieß et al. (2011) for voice alternation in German. Another possibility to derive packaging preferences from salience metrics is to train a classifier that makes use of cumulated salience scores as one (or even the only) factor.

3 Parameters of salience metrics

The framework sketched above specifies a number of parameters of salience metrics, i.e.,

- salience factors that involve (a) different aspects of the linguistic realization of previous/subsequent mentions of the referent, (b) different distance measurements from the last mention of the referent, or (c) different frequency measurements,

- weights of salience factors for the calculation of $hsal(r)$ and $ssal(r)$,

- weights of $hsal(r)$ and $ssal(r)$ for the calculation of cumulated salience scores, and
optional realization thresholds

Different theories of referential coherence entail different parameter configurations, as observed by Hajic'ova and Krujff-Korbayova (1997), Krahmer and Theune (2002) and others for differences between Centering and FGD, and by Poesio et al. (2004) for different instantiations of Centering. The parameter configurations for these theories, as well as for Givon’s Topicality – whose operationalization as part of an NLG system has not been considered so far – are shortly introduced below.

### 3.1 Topicality parameters

Givon (1983, 2001) established two dimensions of ‘topicality’ – abbreviated TOP –, anaphoric topicality and cataphoric topicality, and described correlations between both dimensions of topicality and the choice of grammatical devices.

The anaphoric topicality of a referent $r$ is measured by the distance from its last mention, cataphoric topicality by its persistence (frequency) within the subsequent $n$ utterances:

$$\text{dist}_{cl}(r) = \begin{cases} \frac{1}{k+1} & \text{with } k \geq 0 \text{ intermediate clauses since last mention of } r \\ 0 & \text{if no previous mention of } r \end{cases}$$

$$\text{persist}_{n/cl}(r) = \frac{\text{mentions of } r \text{ within the next } n \text{ clauses}}{n}$$

Here and below, the subscript $cl$ indicates that a factor is defined with reference to clauses. Alternatively, sentences could be considered (subscript $s$).

Hearer salience corresponds to anaphoric topicality, and speaker salience to cataphoric topicality, i.e., $\text{hsal}_{TOP}(r) = \text{dist}_{cl}(r)$ and $\text{ssal}_{TOP}(r) = \text{persist}_{10/cl}(r)$. As for cumulated salience scores, Givon (2001) predicts that (i) high values of $\text{hsal}(r)$ result in high $\text{ref}(r)$ scores (anaphoric topicality favors pronominal realization), and that (ii) high $\text{ssal}(r)$ scores result in high scores for $\text{gr}(r)$ (subject assignment indicates foregrounding):

$$\text{ref}_{TOP}(r) = \text{hsal}_{TOP}(r)$$

$$\text{gr}_{TOP}(r) = \text{ssal}_{TOP}(r)$$

### 3.2 Centering parameters

For Centering (Grosz et al., 1995) – abbreviated CT –, hearer salience corresponds to the ranking of referents in the preceding utterance, with the ranking subject $> \text{object} > \text{other}$, implemented here as an extension of the $\text{dist}(r)$ function above:

$$\text{gr}_{cl}(r) = \begin{cases} \frac{\text{gr}_{ante}(r)}{\text{gr}_{ante}(r) + 1} & \text{with } k \geq 0 \text{ intermediate clauses since last mention of } r \\ 0 & \text{if no previous mention of } r \end{cases}$$

with $\text{gr}_{ante}(r) = \begin{cases} 1.2 & \text{if antecedent is subject} \\ 1.0 & \text{if antecedent is object} \\ 0.8 & \text{otherwise} \end{cases}$

The numerical scores of $\text{gr}_{ante}(r)$ reflect the relative ranking proposed by the theory, and that they are equally distributed between 0.8 and 1.2.7

In accordance with the concept of “backward-looking center” ($C_B$), speaker salience can be defined with respect to the following utterance: A referent is speaker salient if it represents the $C_B$ of the following utterance. To prevent cyclic definitions, the $C_B$ of the following utterance (clause) can

---

7While later studies may involve empirically justified numbers for $\text{gr}_{ante}(r)$, this paper only considers theory-internal evidence to motivate numerical salience factors. The numerical values are thus chosen such that they reflect the original ranking, but the exact numerical values of salience factors are arbitrary. Important for their appropriate interpretation and for the training of decision trees on individual factors is only that relative differences are preserved.
be heuristically identified by pronominal realization (Centering Rule 1):

$$\text{pron}_{\text{ana/cl}}(r) = \begin{cases} 1.0 & \text{iff } r \text{ realized as pronoun in the following clause} \\ 0 & \text{otherwise} \end{cases}$$

Pronominalization is associated with the $C_B$ (Centering Rule 1), i.e., the most (hearer-) salient referent in the current utterance, so that high $\text{ssal}(r)$ scores thus entail high $\text{ref}(r)$ scores:

$$\text{ref}_{\text{CT}}(r) = \text{hsal}_{\text{CT}}(r)$$

Grammatical roles determine the $C_B$ of the following utterance, so that high $\text{ssal}(r)$ scores entail high $\text{gr}(r)$ scores. Further, Centering Rule 2 predicts a preference for $C_B$ continuity, so that $\text{hsal}(r)$ has a positive influence on $\text{gr}(r)$:

$$\text{gr}_{\text{CT}}(r) = 0.5 \times \text{hsal}_{\text{CT}}(r) + 0.5 \times \text{ssal}_{\text{CT}}(r)$$

### 3.3 Functional parameters

Functional Centering (Strube and Hahn, 1999) and Functional Generative Description (Sgall et al., 1986) introduce $\text{hsal}(r)$ factors that evaluate the type of referring expression of the antecedent and its word order: Following Strube and Hahn (1999) the functions $\text{ref}_{\text{top}}^{\text{ante}}(r)$ and $\text{wo}_{\text{top}}^{\text{ante}}(r)$ can be defined in analogy with $\text{gr}_{\text{d}}(r)$ above with the following sub-functions:

$$\text{ref}_{\text{ante}}^{\text{top}}(r) = \begin{cases} 1.2 & \text{iff } r \text{ realized as pronoun, proper name, or simple definite NP} \\ 1.0 & \text{iff } r \text{ realized as possessive NP or complex definite NP} \\ 0.8 & \text{iff } r \text{ realized as indefinite NP} \end{cases}$$

$$\text{wo}_{\text{ante}}^{\text{top}}(r) = (0.8 + 0.4 \frac{m-n}{m})$$

with $m$ number of words in antecedent sentence, and $n < m$ number of words preceding the antecedent.

The functions $\text{ref}_{\text{ante}}^{\text{top}}(r)$ and $\text{wo}_{\text{ante}}^{\text{top}}(r)$ formalize the claim that referents with topical (given) antecedents are more hearer salient than referents with focal (new) antecedents. The opposite claim, formulated by Sgall et al. (1986), requires alternative formulations of these salience factors $\text{ref}_{\text{ante}}^{\text{foc}}(r) := 2 - \text{ref}_{\text{ante}}^{\text{top}}(r)$ and $\text{wo}_{\text{ante}}^{\text{foc}}(r) := 2 - \text{wo}_{\text{ante}}^{\text{top}}(r)$.

### 4 Evaluation

The parameters identified above are evaluated against referring expressions and grammatical roles in two German newspaper corpora that combine syntactic and anaphoric annotations, i.e., a coreference-annotated subcorpus of the NEGRA corpus (Skut et al., 1997; Schiehlen, 2004), and the Potsdam Commentary Corpus (Stede, 2004; Krasavina and Chiarosc, 2007, PCC).

#### 4.1 Pronominalization and hsal metrics

Hearer salience is evaluated with respect to pronominalization. As shown in Fig. 2, personal pronouns are characterized by a high degree of hearer salience (otherwise, a definite description would have been used) and a low degree of speaker salience (otherwise, a demonstrative pronoun would have been used). As speaker salience is neutralized, pronominalization provides a test case for metrics of hearer salience.

For the study of hearer salience, we applied CART and C4.5 decision trees and classified hearer salience scores against the pronominal and nominal realization of third-person referents. Both learning algorithms produced almost identical results (Tab. 1). All hsal factors outperformed the baseline (predict nominal), and with the exception of $\text{dist}_{\text{cl}}(r)$ on NEGRA, this improvement was statistically significant as confirmed by a $\chi^2$ test. For all factors, high salience scores were identified with a preference to pronominal realization, thereby confirming the predicted influence of hearer salience on the choice of referring expressions (Fig. 2).

With respect to plain distance measurements, sentence-level segmentation outperformed clause-level segmentation. This configuration was thus adopted for hearer salience factors that take the form of the antecedent into consideration. The overall best results were achieved with $\text{ref}_{\text{ante}}^{\text{d}}(r)$ and $\text{ref}_{\text{ante}}^{\text{foc}}(r)$.

Closer inspection of the classifier revealed that prominent realization compensates distance, i.e., a referent that is realized in a prominent way in $U_{k-2}$ (e.g., as subject) is more likely to occur as a pronoun than a referent that is realized in a non-prominent way in $U_{k-1}$ (e.g., as non-argument). The classification results did thus not provide a concrete pronom-
Table 1: Correctness of hsal factors for the prediction of nominal and pronominal realization (C4.5), \( \chi^2 \) significance of correctness improvements over baseline

<table>
<thead>
<tr>
<th>salience factor</th>
<th>correctness (significance)</th>
<th>baseline</th>
<th>NEGRA</th>
<th>PCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>dist_s(r)</td>
<td>.819 (not sig.)</td>
<td>.836</td>
<td>p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>dist_s(r)</td>
<td>.845 (p &lt; .001)</td>
<td>.853</td>
<td>p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>gr_s(r)</td>
<td>.845 (p &lt; .001)</td>
<td>.861</td>
<td>p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>ref^{top}_s(r)</td>
<td>.969 (p &lt; .001)</td>
<td>.942</td>
<td>p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>ref^{loc}_s(r)</td>
<td>.969 (p &lt; .001)</td>
<td>.942</td>
<td>p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>wo^{top}_s(r)</td>
<td>.863 (p &lt; .001)</td>
<td>.887</td>
<td>p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>wo^{loc}_s(r)</td>
<td>.861 (p &lt; .001)</td>
<td>.886</td>
<td>p &lt; .001</td>
<td></td>
</tr>
<tr>
<td>total (# ref.exp)</td>
<td>976</td>
<td>2355</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Pronominalization thresholds for ref^{top}_s(r), ref^{loc}_s(r), and gr(r) as identified with a single conjunctive rule learner

<table>
<thead>
<tr>
<th>salience factor</th>
<th>corpus</th>
<th>threshold</th>
<th>prec</th>
<th>recall</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>gr_s(r)</td>
<td>PCC</td>
<td>.472</td>
<td>.695</td>
<td>.84</td>
<td>.761</td>
</tr>
<tr>
<td>gr_s(r)</td>
<td>NEGRA</td>
<td>.472</td>
<td>.569</td>
<td>.837</td>
<td>.678</td>
</tr>
<tr>
<td>ref^{top}_s(r)</td>
<td>PCC</td>
<td>.523</td>
<td>.830</td>
<td>.899</td>
<td>.863</td>
</tr>
<tr>
<td>ref^{top}_s(r)</td>
<td>NEGRA</td>
<td>.523</td>
<td>.782</td>
<td>.913</td>
<td>.842</td>
</tr>
<tr>
<td>ref^{loc}_s(r)</td>
<td>PCC</td>
<td>.389</td>
<td>.631</td>
<td>.899</td>
<td>.741</td>
</tr>
<tr>
<td>ref^{loc}_s(r)</td>
<td>NEGRA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In experiments with a single conjunctive rule learner (that forces a binary partition of salience scores) ref^{top}_s(r) outperformed the other factors in precision and recall of pronoun prediction (Tab. 2). For subsequent experiments, we adopt ref^{top}_s(r) as the primary metric of hearer salience.

4.2 Subject role assignment and ssal metrics

Speaker salience is evaluated here against the assignment of grammatical roles. The subject represents either a high degree of hearer salience or a high degree of speaker salience (Fig. 3). For the study of speaker salience, we eliminated the influence of hearer salience by considering only sentences where one non-subject referent was at least as hearer salient (ref^{top}_s(r)) as the subject. The relatively low number of sentences that match this pattern (approx. 10%) indicates that subjects tend to be hearer salient. To compensate for data sparsity, data from NEGRA and PCC was combined.

We trained decision trees to predict subject or non-subject realization (Tab. 3). Both C4.5 and CART classifiers confirmed that high speaker salience entails a subject preference.

All persistence measurements outperform the baseline (non-subject), and we find that sentence-level segmentation performs better than clause-level segmentation. As for Centering-inspired speaker salience factors that address the pronominalization of the anaphor, three different variants were tested: pronominalization in the immediately following clause pron_{ana/cl}(r), in the immediately following sentence pron_{ana/s}(r) and pronominalization of the anaphor without contextual restriction pron_{ana}(r). Factor pron_{ana}(r) achieved highest correctness, closely followed by pron_{ana/s}(r), and persist_{1/s}(r) and then by persist_{10/s}(r). For other salience factors, the correctness improvement over the baseline was marginally significant or insignificant.

4.3 Beyond pronouns and subjects

Having identified ref^{top}_s(r) and pron_{ana}(r) as suitable measurements of hearer salience and speaker salience, Fig. 4 illustrates their application to NP complexity and grammatical roles. Different grammatical devices are ordered according to their average salience scores. Edges between two scores indicate highly significant differences between the
Figure 4: Average salience scores for selected grammatical devices (NEGRA+PCC)

<table>
<thead>
<tr>
<th>ref_{ana}(r)</th>
<th>pron_{ana}(r)</th>
<th>referring expressions + length (in words)</th>
<th>grammatical roles</th>
</tr>
</thead>
<tbody>
<tr>
<td>.082</td>
<td>.351</td>
<td>.189</td>
<td>.119</td>
</tr>
<tr>
<td>.166</td>
<td>.189</td>
<td>.142</td>
<td>.266</td>
</tr>
<tr>
<td>.209</td>
<td>.142</td>
<td>.181</td>
<td>.360</td>
</tr>
<tr>
<td>.272</td>
<td>.181</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.336</td>
<td>.373</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.372</td>
<td>.372</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results obtained mirror the theory-based predictions on salience metrics summarized in Figs. 2 and 3. Remarkable here is that these phenomena were not taken in consideration when the salience metric was developed (resp., a salience factor selected for its approximation). For pron_{ana}(r) and ref_{ana}(r), these effects were not even anticipated by the researchers who proposed the salience factors in the first place: Neither Centering nor Functional Centering predict a difference between complex and non-complex proper names. Such differences are, however, fully in line with assumptions of the theoretical literature, Ariel (1990), for example, postulated a gradual decrease of complexity with increasing salience.

Figure 4 shows two types of extensions in the application of salience metrics as compared to the data sets they were developed on: (1) change of domain (pron_{ana}(r) applied to referring expressions), and (2) change of granularity (pron_{ana}(r) applied to differentiate non-subject referents, ref_{ana}(r) applied to differentiate nominal expressions). For both types of extension, the theory-based predictions of the MSF could be confirmed, and on this basis, a classifier for packaging preferences can be developed (Sect. 2.5). For the development of such a classifier from an establishes salience metric, it is sufficient to consider only the salience scores and the respective target realizations. With so few parameters, a small amount of data is sufficient to train a classifier for this task.

This is of practical relevance to NLG because it allows us to develop a salience metric for an easily observable phenomenon with loads of training data, and then apply it to another domain, where little training data is available, just sufficient to perform the necessary adjustments (e.g., to calculate the relative weight of hearer salience and speaker salience for the phenomenon under discussion). An interesting prediction is, for example, that speaker salience (and absence of hearer salience) entails differences in accentuation (following Ariel, 1990, and Levelt, 1989, prosodically prominent expressions are more ‘complex’ than prosodically non-prominent expressions, and thus subject to the complexity predictions of Fig. 2). Corpora with prosodic and coreference annotation are available, but expensive to create, and thus relatively small (e.g., the German radio news corpus DIRNDL, with 3221 sentences annotated for prosody and information structure, Eckert et al., 2011). But with salience metrics developed for text corpora, this limited amount of data is sufficient to evaluate whether the salience metrics yield the predicted effects, and to develop a classifier for the salience-based prediction of prosody from previously established metrics.

5 Results and Discussion

This paper described the application of a framework of salience in discourse that introduces a formal distinction between metrics of (backward-looking) hearer salience and (forward-looking) speaker salience, and a definition of information packaging as an alignment between the salience ranking of discourse referents and hierarchies of grammatical devices.

Our model extends Centering in that it assigns every referent a numerical score rather than concentrating on the top-level element in a ranking of ref-
erents from the preceding utterance. By doing so, it is possible to study the effect of distance measurements and to predict packaging preferences for all referents in an utterance, whereas Centering is restricted to adjacent utterances and constraints on possible realizations of the backward-looking center and the preferred center only. Further, our framework is not restricted to pronominalization, but capable to cover elaborate hierarchies of referring expressions.

Evaluation results on the choice of referring expressions and grammatical roles in German confirmed the theoretical predictions on how hearer salience and speaker salience affects both packaging phenomena (cf. Figs. 2 and 3). Essential assumptions about packaging hierarchies and associated aspects of salience could thus be confirmed.

(Subhierarchies of) the rankings in Figs. 2 and 3 have previously been applied in NLG: Fig. 2 covers standard assumptions about pronominal, definite and indefinite descriptions that can be found in similar form in the GRE algorithms of Dale and Reiter (1995) and McCoy and Strube (1999), and in the generation direction of optimality-theoretic models of anaphor interpretation and generation (Beaver, 2004; Byron and Gegg-Harrison, 2004). Thus salience ranking of grammatical roles has been employed for lexicalization of verbs, e.g., by Stede (1998). Zarrieß et al. (2011) describe an experiment to generate voice alternation on the basis of an implicit notion of hearer salience (‘information status’, approximated from surface features such as pronominalization and definiteness, cf. (Cahill and Riester, 2009) for a similar approach on word order).

The two-dimensional model of salience generalizes over Centering, Topicality and FGD, but it also allows us to formulate novel predictions, e.g., that subsequent pronominalization has an effect on NP complexity, or that the same notion of speaker salience is affecting both grammatical roles and the choice of referring expressions. Both claims have not been stated as such within the original theories.

Furthermore, the evaluation showed that the theory-guided adaption of salience metrics from one packaging phenomenon to another is possible. The theoretical background model adopted here may thus provide us with an opportunity to develop salience-based predictors for domains with relative little training data available.

By combining information drawn from different packaging phenomena, new metrics of salience may be developed and integrated into existing NLG algorithms to predict referring expressions and grammatical roles (as well as word order) in a contextually adequate way.

Acknowledgements

The research described in this paper was partially conducted within the Collaborative Research Center (SFB) 632 ‘Information Structure’, Universität Potsdam, and financially supported by the Linguistics Department, Universität Potsdam. I would also like to thank Manfred Stede and three anonymous reviewers.

References


Linguistic Annotation Workshop. Held in Conjunction with the ACL-2007, pages 156–163, Prague, Czech Republic, June.


