Capturing attention

Determining information structure objectively
A research proposal
Erwin R. Komen
Version 4.6

Section 2a (*max 2000 words, currently 1956 + 42 woorden in Figure)*

1 Information structure

Writers (and speakers) organise the information in clauses, sentences, and texts in a way that takes into account what the reader already knows (contextually given or topical information) and they direct their attention to those parts of the information they consider to be new or important (focus). This is "information structure," and it interacts with clause structure in the sense that it influences the choice of particular syntactic constructions. Consider the following examples, where the focus domains are underlined:

(1) a. In a hole in the ground there lived a (*the) hobbit. (Tolkien, 1937)
Not a nasty, dirty, wet hole, filled with the ends of worms and an oozy smell, nor yet a dry, bare, sandy hole with nothing in it to sit down on or to eat: it was a hobbit-hole, and that means comfort.
... This hobbit was a very well-to-do hobbit, and his name was Baggins.
b. The (*a) hobbit lived in a hole in the ground.
c. It's in a hole in the ground that a hobbit lives.
d. Hobbit Bilbo lived in a hole in the ground.

Although all sentences contain the same basic elements, (i) a location *hole in the ground*, (ii) an action *live* and (iii) a participant *hobbit*, and have the same propositional content, the information is organized in different ways. Sentence (1a), syntactically a presentational construction, introduces a new and major participant (a *hobbit*, who is after three quarters of a page picked up as a main topic by *this hobbit*) against the background of a location (a *hole in the ground*). The regular and neutral word order construction in (1b) is a comment about the topical subject *the hobbit*, which must have been mentioned earlier in the text, and the *it*-cleft in (1c) focuses on one constituent. Each of these different organizations is only valid at a particular point in a text, where it is of crucial importance for proper understanding: (1a) introduces a new major participant, (1b) refers back to an established participant, and (1c) corrects a misunderstanding about a hobbit's habitat.

Information structure bridges the clause-level and the text-level. At the clause-level, it assigns constituents to different components of informational structure (e.g. focus domain, topic, point-of-departure). The link to the text-level is formed by the referential status of the constituents, such as the subject "Hobbit Bilbo" in (1d); if this subject has an antecedent in the preceding context, the information structure becomes "topic-comment", as in (1b), but if the subject is totally new, the focus domain includes the subject, as in (1a).

The knowledge needed by language users for a felicitous use of information structure is of a particularly subtle kind, involving the interaction of grammar, syntax, pragmatics and context-dependent marking of referential state. It typically requires (near) native command of a language, and this is why it is particularly problematic in second language learning, and in a variety of contexts where information structure needs to be 'imitated', such as automatic translation.

1.1 Major claim

My main claim is that information structure is compositional: it derives from syntax and referentiality:

(2) *Information structure compositionality*
Syntax + Referentiality → Information Structure

Syntax consists of constituenthood, word order, phrasal categories and so on, and is restricted to the clause; referentiality, as defined in Komen (2013), consists of referential states (restricted to a set of five primitive categories, the 'Pentaset': New, Inferred, Identity, Assumed and Inert) and, where available, a link to an antecedent. A particular syntactic construction combined with referentiality values for its constituents, leads to only one information structure. This is the claim I take as a basis in this project.

1.2 Challenges

The major challenges in the emerging field of information structure are both theoretical and practical. The theoretical problem is that referentiality and information structure tend to be mixed, but need to be kept apart.

(3) a. (*context:* He (=Bilbo) had only just had breakfast, but he thought a cake or two and a drink of something would do him good after his fright.)
Gandalf in the meantime was still standing outside the door.
[Identity] [Inferred] [Identity] -- *referential status*
[Topic] [Spacer] [Comment] -- *inform. structure*
b. After a while he stepped up.
[New] [Identity] -- *referential status*
[Departure] [Topic] [Comment] -- *information structure*

The referential states (Identity, Inferred, Identity) of the constituents in (3a) need to be combined with the syntax (a referentially old subject and a verb phrase with a referentially old constituent are separated by an adverbial of time) in order to understand the information structure as signalling "topic-switch" (a topic-comment is divided by a spacer, and the topic is relatively older then the topic of the previous sentence). The referential states of the constituents in (3b), likewise, are not informative by themselves. The referentially new information *a while* is part of a typical clause-initial point of departure, which is then followed by a common topic-comment structure.

My compositionality theory regards 'referentiality' as a lower-level primitive, while it regards information structure as being derived, so of a 'higher order'. It is, I claim, the theoretical confusion between the two concepts that lies at the root of the practical problems: direct manual annotation in terms of information structure is often done with categories such as "aboutness topic", "framesetting topic" and different blends of "focus" (Dipper et al., 2004, Zeldes et al., 2009), and typically results in poor interrater agreement (Cook and Bildhauer, 2012). Successful annotation categorizes noun phrases in terms of referentiality, but does not automatically yield information structure (Ariel, 1999, Baumann and Riester, 2013, Gundel et al., 1993, Haug et al., 2009, Lambrecht, 1994, Prince, 1981, Riester et al., 2010).

I have developed a software program "Cesax" that allows semi-automatic labelling of referentiality, and it is the combination of having a small set of only five categories in combination with the semi-automatic process that results in a high interrater agreement (Komen, 2011a). My project will use the outcome of this successful annotation scheme to calculate the higher level information structure objectively.

2 Innovativeness

Three elements in my project are innovative: (i) the theory I propose, (ii) the algorithm I develop, (iii) the text-oriented slant I take, which leads me to develop new methodology.

As for (i) and (ii), I intend to develop the theoretical underpinnings for my claim in (2), and the crucial component of formulating the rules linking syntax and referentiality to information structure in the shape of an algorithm that automatically derives the information

structure of natural texts. These texts need to contain the necessary syntactic and referential information (where they don't, I will further develop and use semi-automatic referentiality annotation to add this).

As for (iii), my approach to calculate information structure automatically will be text-oriented. The primary reason for this is that information structure bridges the clause-level and the text-level, hence information structure calculation should include the text-level. Secondly, texts provide structures that are homogenous in terms of authorship, date and genre. I intend to include the text-level by developing a method that evaluates texts as a whole, in order to assign clausal constituents to the appropriate columns in a tabular representation of these texts. My method will result in *Optimal Tabular Text Representations* ("OTTERs").

3 Methodology

Calculating OTTERs is my own novel method, and is inspired by the existing method of 'charting' texts, which maps the major constituents in each clause for the purpose of discourse analysis (Dooley and Levinsohn, 2001, Grimes, 1975, Huttar, 2003, Longacre and Joo, 2012, Macleod, 2003, Ulfers, 1993). Tabular representations show the relation between the clause-oriented syntax and the structure of the text:

Table 1 Tabular text representation (data from Tolkien 1937)

	Conj	Pre	Sbj	Vb	Obj	Post
1		In a hole in the ground	there	lived	a hobbit [Postposed Sbj]	
2a			—	—	Not a nasty, dirty, wet hole, filled with the ends of worms and an oozy smell, nor yet a dry, bare, sandy hole with nothing in it to sit down on or to eat:	
2b			it_1	was	a hobbit-hole	
2c	and		that	means	comfort	
8a			This $hobbit_2$	was	a very well-to-do hobbit,	
8b	and		his name	was	Baggins.	
9a			The Bagginses	have lived	in the neighbourhood of The Hill	for time out of mind,
9b	and		people	considered	them	very respectible,
9c		not only because	most of them	were	rich,	
9d	but	also because	they	never had	any adventures	
9e	or		—	did	anything unexpected:	
9f			you	could tell	what a Baggins would say on any question	without the bother of asking him.

The table shows the outliers against the background of the most common topic-comment information structure. This provides an important pre-processing step for the algorithm.

I intend to achieve the necessary *objectivity* in OTTERs by developing a method that calculates how the constituents in clauses most optimally 'fit' the columns in a table, in the sense that the result has the best combination of specificity (one column for one type of constituent), entropy (diversity in constituent types per column), sparseness (empty cells) and stacking (cells with more than one constituent).

The algorithm that calculates information structure makes use of already available syntactic annotation, the OTTERs, and the referentiality annotation, as illustrated in Figure 1.

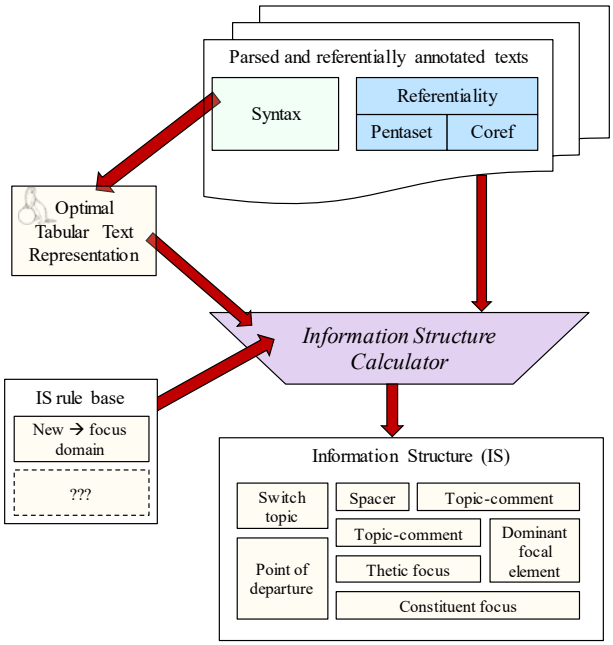

Figure 1 Using OTTERs and referentiality to calculate information structure

The development of my approach initially makes use of four syntactically annotated historical corpora that can be regarded as natural text (Kroch et al., 2004, Kroch et al., 2010, Kroch and Taylor, 2000, Taylor et al., 2003). Eighteen of these have been referentially enriched by me and my colleagues, as indicated in Figure 1, and they form the basis for developing OTTERs as well as the information structure calculation algorithm. The development cycle is expected to result in rules that determine how syntax and referentiality result in information structure.

Since my claim about the compositionality of information structure is not tied to one particular language, part of my project will be devoted to include other languages. Three languages that I intend to include are Dutch, German and Chechen. The Dutch and German languages help gain insight into historical English, while Chechen provides an additional challenge for the method I propose, given its non-Indo-European syntax and its lack of definite/indefinite articles.

The extensions into other languages, as well as the checking and improving of the available English texts, require referentiality annotation that is solid, trustworthy and speedy. I will do this semi-automatically with the available program “Cesax”, which has shown its value in the annotation of English texts (Komen, 2011b).

4 Timetable and work plan

	GY1	GY2	GY3	GY4
Activity				
• Improving referentiality annotation to gold standard				
• Adding annotated texts				
• Extending and verifying "IS rule" base				
• Developing OTTER algorithm + software				
• Developing IS *calculator*				
• Testing IS calculator on English narrative				
• Workshop to inspire using IS calculation				
• Testing IS calculator on other genres in English				
• Developing referentiality predictor				
• Extending software to include other languages (Dutch, German, Chechen)				

5 Collaboration

The network of researchers I know and can approach for questions and exchange of ideas includes, first of all, scientists from the Radboud University Nijmegen, where I am currently employed: Ans van Kemenade (historical English), Olaf Koeneman (present-day English), Antal van den Bosch (computational linguistics), Helen de Hoop (semantics) and Pieter Muysken (typology and language acquisition).

I also have colleagues within SIL-international, of which I am a member: Linda Humnick (USA, Caucasian languages), Connie Kutsch Lojenga (Leiden, African languages), and Stephen Levinsohn (UK, discourse and information structure).

My other international contacts include: Johanna Nichols (USA, Caucasian languages), Matti Rissanen (Finland, English corpus work), Bettelou Los (Edinburgh, historical English), Susan Pintzuk & Ann Taylor (UK, historical English), Dag Haug (Norway, referentiality annotation), Susanne Winkler (Germany, information structure).

6 Bibliography

Ariel, Mira. 1999. *Accessing noun-phrase antecedents*. London and New York: Routledge.

Baumann, Stefan, and Arndt Riester. 2013. “Coreference, lexical givenness and prosody in German”. *Lingua* 11:16-37.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. “Predicting the dative alternation”. *Cognitive Foundations of Interpretation*. ed. by Gerlof Bouma, Ineke Kraemer and Joost Zwarts, 69-94. Amsterdam, Royal Netherlands academy of science.

Čmejrek, Martin, Jan Cuřín, and Jiří Havelka. 2003. “Czech-English dependency-based machine translation”. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics (EACL '03)* 83-90. Stroudsburg, PA, USA, Association for Computational Linguistics.

Comrie, Bernard. 1989. *Language universals and linguistic typology*. Chicago: The university of Chicago press.

Cook, Philippa, and Felix Bildhauer. 2012. “Identifying "aboutness topics": two annotation experiments”. *Dialogue & Discourse* 3:118-141.

Dipper, Stefanie, Michael Götze, Manfred Stede, and Tillmann Wegst. 2004. "ANNIS: a linguistic database for exploring information structure". *Working Papers of the SFB 632, Interdisciplinary Studies on Information Structure (ISIS)*. ed. by Shinichiro Ishihara, Michaela Schmitz and Anne Schwarz, 245-279: <http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/10023 >.

Doherty, Monika. 1997. "Textual garden paths - parametrized obstacles to target language adequate translations". *Machine translation and translation theory*. ed. by Christa Hauenschild and Susanne Heizmann, 69-90. Berlin, Mouton de Gruyter.

Dooley, Robert A., and Stephen H. Levinsohn. 2001. *Analyzing discourse: basic concepts*: Summer Institute of Linguistics.

Firbas, Jan. 1964. "From comparative word-order studies". *BRNO studies in English* 4:111-126.

Grimes, Joseph Evans. 1975. *The thread of discourse*. The Hague: Mouton.

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

Haug, Dag T. T., Marius L. Jøhndal, Hanne M. Eckhoff, Eirik Welo, Mari J. B. Hertzenberg, and Angelika Müth. 2009. "Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages". *TAL* 50:17-45. <http://www.atala.org/IMG/pdf/TAL-2009-50-2-01-Haug.pdf >.

Huttar, Lars Andrew. 2003. *Constituent Charting for Discourse Analysis: Information Model and Presentation Model*. MA thesis, Graduate Institute of Applied Linguistics

Kaiser, Elsi, and John C. Trueswell. 2004. "The Role of Discourse Context in the Processing of a Flexible Word-Order Language". *Cognition* 94:113-147.

Komen, Erwin R. 2007. *Focus in Chechen*. Master's Thesis, Leiden University

Komen, Erwin R. 2011a. *Cesax: coreference editor for syntactically annotated XML corpora* Nijmegen, Netherlands: Radboud University Nijmegen, <http://erwinkomen.ruhosting.nl/software/Cesax >.

Komen, Erwin R. 2011b. *Cesax: coreference editor for syntactically annotated XML corpora. Reference manual* Nijmegen, Netherlands: Radboud University Nijmegen, <http://erwinkomen.ruhosting.nl/software/Cesax/Cesax_Manual.pdf >.

Komen, Erwin R. 2013. *Finding focus: a study of the historical development of focus in English*. Utrecht: LOT.

Kroch, Anthony, Beatrice Santorini, and Ariel Diertani. 2004. *Penn-Helsinki Parsed Corpus of Early Modern English*, <http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html >.

Kroch, Anthony, Beatrice Santorini, and Ariel Diertani. 2010. *Penn parsed corpus of modern British English*, <http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html >.

Kroch, Anthony, and Ann Taylor. 2000. *Penn-Helsinki Parsed Corpus of Middle English, second edition.*, <http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/ >.

Lambrecht, Knud. 1994. *Information structure and sentence form. Topic, focus and the mental representations of discourse referents.*: Cambridge university press.

Longacre, Robert E., and Hwang Shin Ja Joo. 2012. *Holistic discourse analysis*. Dallas, Tex.: SIL International.

Macleod, Catherine. 2003. *Reference in Udi narrtive discourse*, University of North Dakota

Prince, Ellen. 1981. "Toward a taxonomy of given-new information". *Radical Pragmatics*. ed. by Peter Cole, 223-255. New York, Academic Press.

Riester, Arndt, David Lorenz, and Nina Seemann. 2010. A recursive annotation scheme for referential information status. In *Proceedings of the seventh international conference of language resources and evaluation (LREC)*, 717-722. Valletta, Malta.

Tavecchio, Lotte. 2010. *Sentence patterns in English and Dutch: a contrastive corpus analysis*. Ph. D. dissertation, LOT

Taylor, Ann, and Susan Pintzuk. 2012. "Rethinking the OV/VO alternation in Old English: The effect of complexity, grammatical weight, and information status". *The Oxford handbook of the history of English*. ed. by Terttu Nevalainen and Elizabeth Closs Traugott, 835-845, Oxford university press.

Taylor, Ann, Athony Warner, Susan Pintzuk, and Frank Beths. 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*, <http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm >.

Tolkien, John Ronald Reuel. 1937. *The Hobbit*: Houghton, Mifflin.

Ulfers, Robert Ernst. 1993. *Narrative genre text collection for discourse grammar analysis in Karang*. Yaoundé, Cameroon: Société Internationale de Linguistique.

van Kemenade, Ans, and Marit Westergaard. 2012. "Syntax and Information Structure: verb second variation in Middle English". *Information Structure and Syntactic Change in the History of English*. 1, ed. by Bettelou Los, María José López-Couso and Anneli Meurman-Solin, 87-118. New York, Oxford University Press.

van Vuuren, Sanne. 2013. "Information structural transfer in advanced Dutch EFL writing: a cross-linguistic longitudinal study". *Linguistics in the Netherlands 2011 [AVT30]*. ed. by Suzanne Aalberse and Anita Auer, 173-187. Amsterdam, John Benjamins.

Verheijen, Lieke, Bettelou Los, and Pieter de Haan. 2013. "Information structure: the final hurdle? The development of syntactic structures in (very) advanced Dutch EFL writing". *Dutch Journal of Applied Linguistics* 2:92-107.

Zeldes, Amir, Julia Ritz, Anke Lüdeling, and Christian Chiarcos. 2009. "ANNIS: A Search Tool for Multi-Layer Annotated Corpora". Paper presented at *Proceedings of Corpus Linguistics 2009*, Liverpool, UK.

7 Knowledge exchange and impact

7.1 Potential

The theoretical and practical developments in the emerging field of 'information structure' that I intend to address in my project have a side range of implications for areas such as translation, where it is one of the key factors determining successful communication, language learning, genre description (see for instance Tavecchio, 2010), author recognition, reading level determination, synchronic linguistics and diachronic linguistics. Further implications are in an important area of computational linguistics that so far seems to have reached its limits: coreference resolution. Since my major claim involves the relation between syntax, referentiality and information structure, I claim that coreference resolution will be improved by taking an estimate of the information structure of a clause into account. A final area where I expect my project to have an impact is that of machine translation. Applying correct information structure of the source language to the translation in a target language is argued to improve machine translations (Čmejrek et al., 2003, Doherty, 1997).

An example where including information structure is of vital importance in translation comes from Chechen. Unlike English, which has a special construction to convey focus (the cleft in (1b)), Chechen word order and morphology for this purpose (Komen, 2007, Komen, 2013). Even contrastive focus can be achieved just by position, as in (4a).[1]

(4) a. Cul sov, q'ooman kuljturan xaznash larjiirash a,
than.that more country-GEN culture-GEN treasures keepers &
t'eqi'uorash a adamash du. [p86-00018:7]
developers & people are

b. #Moreover, those who preserve and develop the treasures of a country's culture are people.

c. Moreover, it is the people who preserve and develop the treasures of a country's culture.

Translating involves detailed knowledge of the information structure of both languages, to avoid a translation sucha as (4b) (where the propositional content is correct, but the information structure is lost), instead of the optimal one in (4c).

It follows that information structure is also important in language learning. First-year Dutch university students majoring in English generally have good command of vocabulary and grammar, but the information structure of their writing is still profoundly infelicitous (van Vuuren, 2013, Verheijen et al., 2013). I expect the theoretical knowledge gained in this project, together with the software, to find an application in monitoring and evaluating the strategies students use to convey information structure in the language they are learning.

In synchronic linguistics, there are word order alternations that can only be fully understood by taking information structure into account. Bresnan et al. (2007), for instance, have shown that the "dative alternation" (*He gave the book to John* versus *He gave John a book*) is sensitive to an information ordering rule: what is relatively old will, where possible, precede what is relatively new (Comrie, 1989, Firbas, 1964, Kaiser and Trueswell, 2004).

As for historical linguistics, there is growing awareness that much of the word order variation recorded is possibly not (only) a matter of change in syntax, but may well be motivated by information structure considerations (Taylor and Pintzuk, 2012, van Kemenade and Westergaard, 2012). The advent of syntactically parsed historical corpora of English (spanning a period from roughly 900 to 1900 A.D.) has facilitated research into the relation

[1] The word order of this copula clause is OSV.

between word order change (such as the use of the VO or OV order in subclauses, and the SV versus VS order in main clauses) and the expression of information structure. Work in these areas requires determining information structure objectively, which is exactly the goal of my project.

7.2 Implementation

The potential knowledge users include linguists studying word order variations, those studying genre characteristics, researchers of second language acquisition, those involved in translation studies (including machine translation), and those working on coreference resolution.

I intend to make the software that I will develop in my project available as *open-source*. This software calculates the information structure of the sentences in a text, provided this text has been syntactically and referentially annotated. The fact that other researchers can make us of this software will promote the dissemination and utilisation of the outcome of this project.

A workshop in the second half of my project aims to inspire researchers to implement the algorithms I provide in their work.