Action Categorisation in Multimodal Instructions

Ielka van der Sluis, Renate Vergeer, Gisela Redeker
Center for Language and Cognition Groningen
{i.f.van.der.sluis, r.vergeer, g.redeker}@rug.nl

Abstract

We present an explorative study for the (semi-)automatic categorisation of actions in Dutch multimodal first aid instructions, where the actions needed to successfully execute the procedure in question are presented verbally and in pictures. We start with the categorisation of verbalised actions and expect that this will later facilitate the identification of those actions in the pictures, which is known to be hard. Comparisons of and user-based experimentation with the verbal and visual representations will allow us to determine the effectiveness of picture-text combinations and will eventually support the automatic generation of multimodal documents. We used Natural Language Processing tools to identify and categorise 2,388 verbs in a corpus of 78 multimodal instructions. We show that the main action structure of an instruction can be retrieved through verb identification using the Alpino parser followed by a manual selection operation. The selected main action verbs were subsequently generalised and categorised with the use of Cornetto, a lexical resource that combines a Dutch Wordnet and a Dutch Reference Lexicon. Results show that these tools are useful but also have limitations which make human intervention essential to guide an accurate categorisation of actions in multimodal instructions.

Keywords: instructions, actions, verbs, categorisation, task structure

1. Introduction

Multimodal instructions (MIs) consist of pictures and text that present a sequence of actions that users of these documents need to carry out to perform a particular procedural task. It is not known exactly which combinations of pictures and text are most effective to present such a procedure in a context of use (Schriver, 1997; Aouladomar, 2005; Bateman, 2014). We advocate the combined use of corpus studies and user studies to determine the effectiveness of picture-text combinations in order to evaluate and (in the future) automatically generate multimodal documents (Van der Sluis et al., 2017). Our previous work on annotation of MI corpora shows how pictures, text and their relations in MIs contribute to the presentation of the actions that need to be carried out to perform a task. For instance, our corpus study in the cooking domain (Van der Sluis et al., 2016b) revealed that even in MIs with text+picture pairs for each step of the procedure, many actions (here 56% of the 452 actions in 30 MIs) are presented only textually and–unsurprisingly–that picture-only presentations are rare (here 7%). When an action is presented in both, text and picture (here 37% of the actions), the information presented is not always the same. This was further explored in a study in the first aid domain (Van der Sluis et al., 2017), where we found that pictures in MIs may present an action-in-progress, but also the result of an action. In the latter case the action itself is described in the text (e.g., 'pick up the tweezers'), while the picture presents the result (depiction of a hand holding tweezers).

In this paper we present our efforts to automatically identify and categorise the actions presented in the text of a corpus of multimodal first aid instructions. Arguably, a categorisation of actions in text will inform the identification of those same actions (and/or their results) in pictures (Ghanimifard and Dobnik, 2017; Vedantam et al., 2015). While automatic identification of depicted objects is well studied, the identification of depicted actions is known to be more difficult (Jensen and Lulla, 1987; Stanfield and Zwaan, 2001; Socher et al., 2014; Karpathy and Fei-Fei, 2015). Ultimately the identification of actions in pictures can help us to determine the type and content of the picture-text relations in multimodal documents. Our research question is formulated as follows: How can the constituent actions in Dutch first aid procedures be identified and categorised by (semi-)automatic natural language analysis, so that the resulting action categories can be used to identify picture-text relations in multimodal instructions?

Automatically acquiring procedural knowledge from instructions without any domain knowledge is challenging (Zhang et al., 2012). Inevitably, some preprocessing should be conducted because human instructions naturally contain imperfections such as ambiguities, omissions and errors. Most of all, discovery of the main action structure requires proper processing of mainly the verbs in instructions (Steehouder and Van der Meij, 2005). Actions in instructions are often represented by imperative verbs because instructions are processed most effectively and efficiently if they are presented in explicit and direct terms (Steehouder and Karreman, 2000; Piwek and Beun, 2001). However, actions can also be represented with the use of a gerund (as in 'Drag the victim out of the danger zone walking backwards.'), which often specifies the way in which the action represented by the main action verb, 'drag', should be carried out. The texts in multimodal instructions also include numerous actions that are not a part of the main action structure, while the accompanying pictures usually visualise some or all of the

main actions in the procedure. The additional actions concern alternatives, contingencies, or cautionary advice, and are verbalised using modal verbs (e.g., 'Small children can just be turned on their side facing downward.'), negations (e.g, 'There should be no pressure on the chest that can make breathing difficult.'), conditionals (e.g., 'If the victim wears glasses, take them off and put them in a safe place.') or warnings (e.g., 'Pay close attention to the victim.'). This paper presents a method to identify and categorise the verbs that represent the main action structure in MI texts as well as the results of this method in a corpus of first aid instructions.

2. Method

2.1. Dataset Preparation

We selected 78 MIs from the annotated PAT corpus (Van der Sluis et al., 2016a). They were published in two editions of Het Oranje Kruis Boekje 2011[1] and 2016[2]. Het Oranje Kruis[3] is a Dutch organisation that provides learning materials for first aid certification trainings. The two editions of Het Oranje Kruis Boekje overlap in terms of the tasks presented in them: 25 tasks appear in both editions (yielding 50 MIs), and 28 appear only in one edition. Preprocessing of the MIs was relatively easy, because the MIs contain fewer imperfections compared to, for instance, MIs published on the internet (often by unauthorised sources). MI texts were augmented by adding periods at the end of every title and every item in enumerated lists to allow automatic identification of sentences. In addition, semicolons in the MI text were changed to colons to avoid confusion with the delimiter used in our data files. The 78 MIs include 1,342 sentences and 2,388 verbs in total.

Figure 1 and Figure 2 present two examples of MIs that display how to place a victim in the recovery position, a life-saving operation to prevent unconscious people to choke in their own fluids. In short, the first aid helper needs to kneel down on one side of the victim, place the victim's legs and arms in particular positions to allow turning the victim on his/her side. Subsequently, the helper has to make sure that the victim's head is placed in such a way that the victim's airway will stay open. Then the victim's breathing has to be checked at regular time intervals. The pictures in both examples display a number of actions in the procedural task that are presented in the MI text (e.g., to place a victim's hand on the victim's face, to bend the victim's knee, to turn the victim on his/her side, to position the victim's head). Moreover, the MI texts also present a number of actions of which the results are visible in multiple pictures (e.g., to kneel, to place an arm of the victim sideways). For instance, in both MIs the first aid helper


De stabiele zijligging vanuit rugligging

1 Kniel naast het slachtoffer aan de kant van zijn gezicht.

2 Als het slachtoffer een bril draagt, neem die dan af.

3 Zorg dat zijn benen gestrekt naast elkaar liggen.

4 Leg de dichtstbijzijnde arm van het slachtoffer loodrecht op zijn lichaam en bij voorkeur met de handpalm naar boven. Let erop dat u niet forceert.

5 Buig de andere arm van het slachtoffer over zijn borst en leg deze met de handrug tegen de wang die het dichtst bij u is. Houd deze hand vast. Let op, duw niet zijn hoofd opzij.

6 Pak het been van het slachtoffer dat het verst van u verwijderd is net boven de knie aan de buitenzijde vast en trek het op terwijl zijn voet op de grond blijft rusten.

7 Draai het slachtoffer naar u toe door aan zijn gebogen knie te trekken. Houd tegelijkertijd de handrug tegen zijn wang geduwd. Draai totdat zijn gebogen elleboog de grond raakt of totdat het slachtoffer niet verder kan draaien.

8 Leg zijn gebogen been zo dat zowel zijn heup als knie aan de buitenzijde een rechte hoek vormen.

9 Kantel zijn hoofd voorzichtig naar achteren en leg zo nodig zijn hand beter onder zijn wang om ervoor te zorgen dat zijn luchtweg vrij blijft. Richt zijn neus en mond naar de grond, zodat vloeistoffen gemakkelijk uit zijn mond kunnen lopen

10 Observeer het slachtoffer continue en voer elke minuut de ademhalingscontrole uit.

5

6

8

Leg bij kinderen zo nodig een klein kussen of opgerolde deken achter de rug om de houding stabiel te maken. Er mag geen druk op de borst zijn die de ademhaling kan bemoeilijken.

Als het slachtoffer zwanger is, moet u haar bij voorkeur op de linkerzij draaien. Op de rechterzij kan het bloed moeilijker terugstromen naar het hart (door druk van het kind op de onderste holle ader).

Figure 1: MI911 Placing a victim in the recovery position (Het Oranje Kruis Boekje, 2013).


Stap 3
Vervolgens buig je de verste arm over de borst van het slachtoffer. Leg deze arm met de handrug tegen zijn wang die het dichtst bij je is. Houd deze hand vast. Let op: duw zijn hoofd niet opzij.

Stap 4
Pak het been dat het verst van je verwijderd is net boven de knie aan de buitenzijde vast. Vervolgens trek je dit been op terwijl zijn voet op de grond blijft rusten.

Stap 5
Draai het slachtoffer naar je toe door aan de gebogen knie te trekken. Houd tegelijkertijd de handrug tegen zijn wang geduwd. Draai door totdat de gebogen elleboog van het slachtoffer de grond raakt of totdat het slachtoffer niet verder kan draaien.

stap 3, 4 en 5

Stap 6
Leg het gebogen been van het slachtoffer zo neer dat zowel zijn heup als zijn knie een rechte hoek vormen. Plaats zo nodig de elleboog beter op de grond.

stap 6

Stap 7
Duw zijn hoofd voorzichtig naar achteren. Leg als dit nodig is zijn hand beter onder zijn wang om ervoor te zorgen dat zijn luchtweg vrij blijft. Richt de neus en mond van het slachtoffer naar de grond, zodat vloeistoffen gemakkelijk uit zijn mond kunnen lopen. Let op dat zijn hand zijn mond niet dichtduwt.

stap 7

Figure 2: Part of MI959 Placing a victim in the recovery position (Het Oranje Kruis Boekje, 2016).

[1]Het Oranje Kruis (2011). Het Oranje Kruis Boekje, De officiële handleiding voor eerste hulp. Thieme Meulenhoff, Amersfoort. ISBN 9789006921717.

[2]Het Oranje Kruis (2016). Het Oranje Kruis Boekje, De officiële handleiding voor eerste hulp. Thieme Meulenhoff, Amersfoort. ISBN13 9789006410341.

[3]http://www.hetoranjekruis.nl/

has already knelt in the first picture and stays on his knees in the second picture. Also the outstretched left arm is visible in multiple pictures in both MIs. We refer to (Van der Sluis et al., 2017), where we also show that in a tick removal instruction the presentation of processes and results of actions differs between the MI the text and the MI pictures. In this paper we present a method to categorise actions in the MI text, to allow the identification of actions and results of actions in MI pictures and to specify text-picture relations in MIs.

2.2. Action Identification

To identify the syntactic segments and their hierarchical relations, the MI texts were processed using the NLTK *sent_tokenizer* (Bird, 2006). Per MI the resulting sentences were parsed with the Alpino parser (Van Noord and others, 2006) and a database was created with the lemmas of the verbs that Alpino identified; for each of the 2,388 verbs in the corpus the database includes the lemmatized verb, the index of the MI and the index of the sentence in the MI in which the verb occurs. The Alpino parser's errors (N=18) were manually resolved through removal of nouns and nominalisations that were mistakenly tagged as verbs (i.e. five times 'buikstoten', three times 'kompressen', twice 'beademingshulpmiddelen' and once 'weerkanten', 'rautekgrijpen', 'bloedhoesten', 'paniekaanvallen', 'sponzen', 'riltklappertanden', 'bevriezingswonden', 'insectsteken').

To exclude modalised, negated, or conditional actions and warnings in the MI texts we consulted the Algemene Nederlandse Spraakkunst (General Dutch Grammar) (Haeseryn et al., 1997)[4] to generate a script to perform a word-based search on the Alpino output. Table 1 presents the features and words identified in the 78 MIs as well as an overview of the 977 verbs (40.9%) that were, as a result, excluded from further analysis: 221 were modal verbs, 567 appeared in the scope of a negation and 514 appeared within a conditional context or as part of a warning. Note that these categories are not mutually exclusive (e.g., 'make sure that you do not strain the arm' contains a warning as well as a negation). We kept 1,411 verbs that represented the main actions in the 78 MIs in our corpus. Subsequently, an overall MI-lemma database was created with 282 unique lemmas for these remaining 1,411 verbs.

2.3. Verb Generalisation

Cornetto (Vossen et al., 2013), a lexical resource that combines a Dutch Wordnet and a Dutch Reference Lexicon, was used to build a verb-hyperonym database that listed all hyperonyms for each lemma in our MI-lemma database. To categorise the verbs in the MI texts, first the synset ID of the verb lemma in the Cornetto 2.0 ID XML database was selected. Subsequently, the hyperonyms in the corresponding synset in the Cornetto 2.0 synset XML database were retrieved. It appeared that for more than a third of the lemmas in the MI-lemma database (N=132) no hyperonyms existed in Cornetto 2.0. In addition, in 71 cases the retrieved Cornetto 2.0 hyperonyms did not fit the meaning of the verbs used in a first aid context. Therefore we manually consulted other sources to find appropriate hyperonyms, i.e. the Cornetto Demo[5] and the Van Dale Dictionary[6]. In the case that none of these sources provided an accurate hyperonym, the verb itself was used as a hyperonym unless the verb contained a prefix. In the latter case the prefix was stripped and the nucleus of the verb was identified as the hyperonym. For example: 'doorduwen' (to push through) became 'duwen' (to push). Table 2 presents the origin of the 92 hyperonyms retrieved for the 282 unique verbs in our MI-lemma database. The 21 hyperonyms selected for further analysis have a frequency $> 1\%$ and do not include semantically weak verbs such as 'zijn' (to be), 'gaan' (to go), 'hebben' (to have), 'komen' (to come). The Cornetto Demo was used to structure the 21 hyperonyms.

3. Results

The 21 hyperonyms subsume numerous verbs in our dataset with only 78 instructive texts. The hyperonym with the highest frequency is 'handelen' (to do, N=155), which fits a corpus with instructions well. The verbs it subsumes vary in frequency: 'laten' (to let, 34.2%), 'doen' (to do, 25,2%), 'zorgen' (to care, 21,9%), 'overnemen' (to pass, 5.8%, 'helpen' (to help, 3.9%), 'herhalen' (to repeat 2.6%), 'uitvoeren' (to execute, 1.9%), 'nemen' (to take,1.3%), 'gedragen' (to behave, 1.3%), 'werken' (to work, 0.6%), 'verzorgen' (take care of, 0.6%) and 'steunen' (to support, 0.6%). Most hyperonyms typically include only two or three frequently used verbs and a few less frequent verbs. For instance, the hyperonym 'plaatsen' (to put, to place, N=99) subsumes 'leggen' (to lay, 48%), 'plaatsen' (to place, 44.4%) and three other verbs that together appear only seven times (to lay down, to apply and to cross). Another example is 'vastmaken' (to attach), which subsumes 'aanleggen' (to fit, 37.7%), 'vastzetten' (to fasten, 23%), 'zetten' (to set, 19.7%) and six other verbs with a maximum frequency of 6.6% (to clasp, to hook, to seize, to tie, to append, to attach). An exception is 'veranderen' (to change, N=57), which subsumes 21 verbs of which 'worden' (to become, 29.8%), 'vouwen' (to fold, 12.3%) and 'koelen' (to cool, 12.3%) are the most frequently used.

The 21 hyperonyms were grouped into eight categories based on their synsets included in the Cornetto Demo: 'handelen' (to do, N=155), 'veranderen' (to change, N=57), 'houden' (to hold, to keep, N=32), 'geven' (to give, N=26), 'voortbewegen' (to propel, N=20), 'onderzoeken' (to investigate, N=20), 'contacteren' (to

[4] http://ans.ruhosting.nl/

[5] http://www.cltl.nl/results/demos/cornetto/

[6] http://www.vandale.nl/

Feature	Words	Excluded Verbs
Modal verbs	kunnen (can, N=161), moeten (should, N=30), mogen (may, N=10)	221 (9.3%)
Explicit negation	niet (not)	297 (12.4%)
Negation with 'niet'	geen (none, N=70), niemand (no one, N=14), nooit (never, N=7), niets (nothing, N=5)	96 (4.0%)
Other negation element	alleen (only, N=49), maar (but, N=30), zonder (without, N=28), minder (less, N=25), enkel (just, N=10), hoogstens (at most, N=4), slechts (only, N=9), nauwelijks (barely, N=7), pas (only just, N=4), weinig (few, N=4), zelden (rare, N=2), moeilijk (hard, N=2)	174 (7.3%)
Conditional	als (if, N=226), wanneer (when, N=170), zolang (as long as, N=8), indien (if, N=8)	408 (17.1%)
Warning	voorkomen (to prevent, N=56), opletten (to pay attention, N=46), uitkijken (to watch out, N=4)	106 (4.4%)
Excluded verbs		977 (40.9%)
Main verbs		1,411 (59.1%)
Total		2,388 (100%)

Table 1: Features and words used to identify modal verbs, negated actions, conditional actions and warnings in the parsed MI texts and the number and percentages of excluded verbs.

Source	Unique Verbs	Total Nr. of Verbs
Cornetto 2.0 DB	127 (45%)	938 (66.5%)
Cornetto Demo	118 (41.8%)	365 (25.3%)
Van Dale	25 (8.9%)	83 (5.9%)
Prefix stripping	12 (4.3%)	25 (1.7%)
Total	282 (100%)	1,411 (100%)

Table 2: Hyperonym sources.

contact, N=18) and 'schoonmaken' (to clean, N=15). Table 3 presents the categories and the hyperonyms included in them, their frequencies, and some examples from our corpus.

4. Discussion

We conclude that the constituent actions in Dutch first aid instructions can be identified by the following procedure: (1) selection of the verbs from MIs, (2) exclusion of modalised actions, negated actions, conditional actions, and warnings, (3) selection of hyperonyms for the remaining verbs and (4) abstraction from hyperonyms to synsets. In this procedure, existing tools to automatically analyse the Dutch MI dataset are helpful, but not sufficient. The research presented in this paper was strongly dependent on natural language processing (NLP) tools created for Dutch. These tools were not entirely complete and reliable. In some cases output was missing, in other cases the output was inappropriate. As a consequence substantial manual support was needed.

Although the Alpino parser is definitely useful to identify verbs, it was unable to retrieve all verbs in our corpus. For instance, Alpino failed to recognise the verb 'inademen' in the instruction 'Adem normaal in en plaats uw wijdgeopende mond goed sluitend over de mond van het slachtoffer' (MI915: Breathe in normally and place your widely opened mouth tightly on the mouth of the victim). Conversely, some words in our corpus were mistakenly tagged as verbs. Sometimes prefixes of separable verbs were not included in the lemma of the verb. For example, Alpino tagged the verb 'plaatsnemen' (to sit down; in MI950: 'Neem plaats achter het slachtoffer' i.e. Sit down behind the victim), as 'nemen' (to take), while the separately occuring prefix 'plaats' is crucial to interpret the meaning of the whole verb. Although in this case manual correction would have been possible, we did not bother with it as our goal was to retrieve hyperonyms. Currently, Alpino does not provide a repair strategy to manually add or replace tags. Because Alpino does not provide information about negations, conditions and warnings, the exclusion of verbs that are not part of the main procedure in the instruction had to be done manually. We chose an approach based on signalling words to discover verbs outside the main procedure of the instruction (see Table 1). Consequently, there is a risk that some main action verbs were incorrectly excluded from further analysis. Other features not included in Alpino that might be useful to determine if a verb describes a main action would be the recognition of causals to identify reasons for doing something, verb tenses and disjunctions.

Cornetto provided hyperonyms for about two thirds of the lemma's in our corpus. The remaining verbs were manually tagged using other sources. Relations between the hyperonyms can be retrieved with the Cornetto Demo. Since one word can have multiple word meanings, it can also have multiple hyperonyms. Because of that, a human annotator is still needed to select the most suitable hyperonym in a particular context. While the selection of hyperonym and synset categories was executed and refined in close discussion be-

	Synset Categories	Freq.	MI-number and Translated Example
1.	handelen (to do)	155 (11%)	903-*Repeat* these last steps until you are out of the danger zone.
2.	veranderen (to change)	57 (4.0%)	987-Someone has immediately *become seriously ill*.
			958-You *take turns* in resuscitating every 2 minutes.
2.1	vastmaken (to attach)	61 (4.3%)	933-*Attach* the bandage with adhesive plaster or a bandage clip.
2.2	bewerken (to manipulate)	28 (2.0%)	959-*Prepare* the breathing mask for use.
2.2.1	dekken (to cover)	17 (1.2%)	971-Then you *cover* the wound with a sterile bandage.
2.3	draaien (to turn)	31 (2.2%)	911-Place your hand on his forehead and *tilt* his head backwards.
2.4	brengen (to bring)	23 (1.6%)	980-*Put* a stifled victim in a half-sitting position and support him.
3.	houden (to hold, to keep)	32 (2.5%)	979-*Hold* his head in the position in which you found it.
			989-*Keep* clinging clothing wet.
4.	geven (to give)	26 (1.8%)	973-This is how you *give* enough support without squeezing.
5.	voortbewegen (to propel)	20 (1.6%)	951-*Slide* both your arms under the victim's armpits.
5.1	verplaatsen (to move)	35 (2.5%)	902-*Lift* him by stretching your legs.
5.1.1	plaatsen (to put, to place)	99 (7.0%)	921-*Place* the CPR face shield on the victim's face.
5.1.2	consumeren (to consume)	28 (2.0%)	988-Give a child something lukewarm with a lot of sugar to *drink*.
5.1.3	verwijderen (to remove)	22 (1.6%)	957-*Remove* any (medicine) plasters from the victim.
5.1.4	trekken (to pull)	25 (1.8%)	935-Carefully *separate* the eyelids with thumb and index finger.
5.1.5	duwen (to push)	38 (2.7%)	905-*Push* the victim on his side.
6.	onderzoeken (to investigate)	20 (1.4%)	914-*Judge* his breathing and start resuscitating.
6.1	waarnemen (to observe)	32 (2.3%)	954-Moreover, the emergency officer on the phone will *hear* you.
6.1.1	zien (to see)	54 (3.8%)	911-Judge his breathing by *looking*, listening and feeling for 10 secs.
7.	contacteren (to contact)	18 (1.3%)	982-Otherwise, *call* the GP's emergency number or *call* the GP center.
8.	schoonmaken (to clean)	15 (1.1%)	972-*Rinse* the victim's eye for 15 minutes with lukewarm water.

Table 3: Eight hyperonyms categories with their synsets, frequencies and corpus examples translated from Dutch to English.

tween the authors of this paper, future research should involve several annotators to allow reliability assessments and improve the validity of the analysis.

5. Future Work

The eight main action categories that we derived from the instructions by Het Oranje Kruis will be tested on the other MIs in our corpus. The growing PAT MI corpus (Van der Sluis et al., 2016a) currently contains 308 MIs with the same topics and tasks included in the materials from Het Oranje Kruis. The PAT corpus will be made available for research purposes when ready. In the future a thoroughly validated categorisation of first aid actions may also be used to recognise first aid actions automatically (with the proviso that manual curation will be needed for a valid and complete coding). Moreover, this categorisation will facilitate parallel coding of the actions presented in the text and in the pictures of the MIs. After the actions have been matched, their textual and pictorial presentations can be compared using (i) more detailed linguistic analyses including aspect, modality, and adverbial specifications of manner, and (ii) more fine-grained visual analysis identifying postures, gaze, and the positions of body parts of the depicted persons. Together with user studies testing the effects of possible pairings, this will eventually facilitate the automatic generation of effective picture-text relations in multimodal documents.

6. Acknowledgements

We thank Johan Bos, Piek Vossen, Isa Maks and Hennie van der Vliet for help and advice, Het Oranje Kruis for advice and the use of the multimodal instructions they published, and the Centre for Digital Humanities at the University of Groningen for partial funding of this project.

7. Bibliographical References

Aouladomar, F. (2005). A semantic analysis of instructional texts. In *Proceedings of the Sixth International Workshop on Computational Semantics (IWCS-6)*.

Bateman, J. (2014). Using multimodal corpora for empirical research. In *The Routledge handbook of multimodal analysis*, pages 238–252. Routledge, London.

Bird, S. (2006). NLTK: the natural language toolkit. In *Proceedings of the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL) on Interactive presentation sessions*, pages 69–72.

Ghanimifard, M. and Dobnik, S. (2017). Learning to compose spatial relations with grounded neural language models. In *Proceedings of the 12th International Conference on Computational Semantics IWCS 2017*.

Haeseryn, W., Romijn, K., Geerts, G., De Rooij, J., and Van den Toorn, M. (1997). *Algemene Nederlandse Spraakkunst [2 banden]*. Groningen/Deurne: Martinus Nijhoff Uitgevers/Wolters Plantyn.

Jensen, J. and Lulla, K. (1987). Introductory digital image processing: A remote sensing perspective. *Geocarto International*, 2(1):65–65.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Piwek, P. and Beun, R.-J. (2001). Relating imperatives to action. In Harry Bunt, editor, *Cooperative Multimodal Communication*, pages 140–155, Berlin, Heidelberg. Springer Berlin Heidelberg.

Schriver, K. (1997). *Dynamics in document design: Creating text for readers*. Wiley, New York.

Socher, R., Karpathy, A., Le, Q., Manning, C., and Ng, A. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Stanfield, R. and Zwaan, R. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, 12(2):153–156. PMID: 11340925.

Steehouder, M. and Karreman, J. (2000). De verwerking van stapsgewijze instructies. *Tijdschrift voor taalbeheersing*, 22(3):220–239.

Steehouder, M. and Van der Meij, H. (2005). Designing and evaluating procedural instructions with the four components model. In *Proceedings of the Professional Communication Conference IPCC 2005*, pages 797–801. IEEE.

Van der Sluis, I., Kloppenburg, L., and Redeker, G. (2016a). PAT Workbench: Annotation and evaluation of text and pictures in multimodal instructions. In Erhard Hinrichs, et al., editors, *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH) at COLING 2016*, pages 131–139.

Van der Sluis, I., Leito, S., and Redeker, G. (2016b). Text-picture relations in cooking instructions. In Harry Bunt, editor, *Proceedings of the Twelfth Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-12) at LREC 2016*, volume 16, pages 22–27.

Van der Sluis, I., Eppinga, A. N., and Redeker, G. (2017). Text-picture relations in multimodal instructions. In Nicholas Asher, et al., editors, *Proceedings of the workshop on Foundations of Situated or Multimodal Communication (FMSC) at IWCS 2017*.

Van Noord, G. et al. (2006). At last parsing is now operational. In *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42.

Vedantam, R., Lin, X., Batra, T., Lawrence Zitnick, C., and Parikh, D. (2015). Learning common sense through visual abstraction. In *The IEEE International Conference on Computer Vision (ICCV)*, December.

Vossen, P., Maks, I., Segers, R., Van Der Vliet, H., Moens, M., Hofmann, K., Tjong Kim Sang, E., and De Rijke, M. (2013). Cornetto: a combinatorial lexical semantic database for Dutch. In *Essential Speech and Language Technology for Dutch*, pages 165–184. Springer.

Zhang, Z., Webster, P., Uren, V., Varga, A., and Ciravegna, F. (2012). Automatically extracting procedural knowledge from instructional texts using natural language processing. In *Proceedings of the Eight International Conference on Language Sources and Evaluation (LREC'12)*, pages 520–527.